THE USE OF THE BENCHMARK DOSE APPROACH IN HEALTH RISK ASSESSMENT

FINAL REPORT

Prepared by

Kenny S. Crump
Bruce C. Allen

Clement International Corporation
1201 Gaines Street
Ruston, Louisiana 71270

and

Elaine M. Faustman

Department of Environmental Health
University of Washington
Seattle, Washington 98195

Prepared for

U.S. Environmental Protection Agency
Risk Assessment Forum

under contract to

Eastern Research Group, Inc.
EPA Contract No. 68-C8-0036

September 1992

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The U.S. Environmental Protection Agency (EPA) frequently employs a reference dose (RfD) or reference concentration (RfC) in setting standards for human exposure to environmental toxicants that are not known to be carcinogenic. An RfD or RfC is a provisional estimate (with uncertainty spanning perhaps an order of magnitude) of a daily exposure (RfD) or continuous inhalation exposure (RfC) to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime (U.S. EPA, 1991). An RfD or RfC is calculated by applying uncertainty factors to the no observed adverse effect level (NOAEL), which represents the highest experimental dose for which no adverse health effects have been documented.

This use of the NOAEL in determining RfDs and RfCs has been criticized by the scientific community (reviewed by Kimmel and Gaylor, 1988) and by the EPA's Science Advisory Board in the course of public review of the Developmental and Reproductive Risk Assessment Guidelines (U.S. EPA, 1986, 1988a, b, 1989) as not making the best use of the available data. These criticisms include the following:

- whether a given experimental dose actually constitutes a NOAEL is subject to scientific judgment and is often a source of controversy;
- experiments involving fewer animals tend to produce larger NOAELs and, as a consequence, larger RfDs or RfCs (the reverse would seem more appropriate in a regulatory context because larger experiments should provide greater evidence of safety);
- the steepness of the dose response plays little role in the determination of the NOAEL; and
- the NOAEL approach does not provide estimates of the potential risks at any exposure levels, in particular those in excess of the RfD or RfC.

Because of these and other limitations of the NOAEL approach for determining RfDs and RfCs, an alternative has been proposed in which uncertainty factors are applied to a benchmark dose (BMD) rather than to a NOAEL (Crump, 1984; Gaylor, 1989). A BMD is a statistical lower confidence limit for a dose that produces a predetermined adverse change in response rate (called the benchmark response or BMR) compared to background. Unlike the NOAEL, the BMD takes into account all of the dose-response information in a study by fitting a mathematical dose-response model to the data. The BMR is generally set near the lower limit of responses that can be measured directly in animal experiments of typical size. Thus, unlike the risk assessment methods that EPA employs with cancer effects (Anderson et al., 1983), the BMD method does not involve extrapolation to doses far below the experimental range.

The BMD approach was proposed as an alternative for determining RfDs and RfCs. It does not share the shortcomings of the NOAEL approach listed above. The BMD approach has other potential advantages over the NOAEL approach that will be discussed later in the document.

This document is a background and guidance document for the application of the BMD approach. The goals, strengths, and limitations of the BMD approach will be discussed, as well as the steps required to implement it. The decisions that must be made at each step are listed, and options for the steps are presented. A detailed comparison is made between the NOAEL and BMD methods. Examples of the steps required in the calculation of BMDs are provided. Finally, areas of additional research related to the BMD approach are suggested.

2. BACKGROUND

2.1. Cancer versus Non-Cancer Effects

Assessment of risk from exposure to toxic chemicals has traditionally been performed differently by EPA depending upon whether the response is cancer or a non-cancer effect (U.S. EPA, 1987). The term non-cancer effect is non-specific and encompasses a wide variety of responses, including adverse effects on specific organs or organ systems, reproductive capacity, viability and structure of developing offspring *in utero*, and survival.[1] For even a single type of effect, the response can range in severity from mild and reversible to irreversible and life-threatening. The severity of the response may depend upon both the level and duration of exposure.

The risk assessment methodology used by EPA for cancer uses dose-response models to extrapolate measured risks to low doses of concern in human populations and for which risks cannot be measured directly. This process of low-dose extrapolation is known to be critically dependent upon the dose-response model selected; different models can fit experimental data equally well, yet yield estimates of risk that differ by many orders of magnitude at low doses (Crump, 1985).

EPA generally uses a dose-response model for estimating cancer risks that assumes that increased risk is proportional to dose at low doses (U.S. EPA, 1987) (i.e., increased risk varies linearly with dose at low doses). An important consequence of this assumption is that any dose, no matter how small, is assumed to result in some increase in risk (i.e., it is assumed that a threshold for response does not exist).

Much of the rationale for these assumptions was based on the idea that carcinogenicity was mediated through genotoxicity. The possibility that a single molecule of a genotoxin may be sufficient to alter DNA in a single cell so that a cancer is eventually produced suggests that—no matter how unlikely such an event is—the dose-response relationship cannot have a threshold and must be linear, at least at low doses (NRC, 1977). Crump et al. (1976) argued more generally that whenever a biological effect occurs spontaneously in the absence of any exposure, and the effect of the toxic insult is mediated through augmenting processes that are already

---

[1]The words "effect" and "response" are used interchangeably in this document and refer generally to conditions that are considered adverse. Although the term "risk" is sometimes used in a similar manner to denote a specific adverse effect (e.g., cancer or reduced fertility), in this document "risk" is used quantitatively and refers specifically to an increased probability of an adverse effect.

operating spontaneously, a threshold would not be present and the response should vary approximately linearly with dose at sufficiently low doses.

In contrast to risk assessment for cancer, dose-response models generally have not been applied by EPA for effects other than cancer. Similarly, less effort has been directed at developing dose-response models for non-cancer effects. One reason for this has been the lack of a consensus regarding the shape of the dose-response curve, especially in the low-dose region, for non-cancer effects. Many scientists believe that thresholds are likely to exist for many chemically induced biological effects, particularly non-cancer effects.

EPA has traditionally set standards based on non-cancer effects by applying uncertainty factors to a NOAEL. This method does not involve use of dose-response models. The purpose of the present document is to discuss an alternative to this approach in which the NOAEL is replaced by a BMD determined using a dose-response model. It is important to keep in mind, however, that the calculation of a BMD does not involve using a dose-response model to extrapolate risks to low doses, as EPA does when conducting risk assessments for cancer effects.

2.2. Overview of the NOAEL Approach to Determining RfDs and RfCs

The BMD and NOAEL approaches have a number of features in common. Before describing the BMD method, a brief description of the NOAEL approach will be presented. A NOAEL has been defined as "that dose of chemical at which there are no statistically or biologically significant increases in frequency or severity of adverse effects between the exposed population and its appropriate control" (Dourson and Stara, 1983). An RfD, or RfC[2], is obtained from a NOAEL by dividing the NOAEL by one or more uncertainty factors.

Different RfDs for the same chemical may be developed for (1) different routes of exposure (e.g., oral RfDs and inhalation RfCs); (2) different durations of exposure (e.g., chronic RfDs for exposures generally lasting from 7 years up to an entire lifetime and subchronic RfDs for exposures generally lasting between 2 weeks and 7 years); and (3) specific types of health effects (e.g., RfDs for developmental effects). The general approach to determining an RfD, which is outlined below, is the same for each type of RfD.

A review of the relevant literature is used to identify the "critical study" upon which the RfD is to be based. This determination takes into account the overall quality of the study, the route and duration of exposure, and range of health effects for which an RfD is desired. If adequate human data are available, such data are used as the basis for the RfD; otherwise data from animal studies are used.

Among the well-conducted studies, the study employing the lowest dose at which a toxic effect is detected is generally selected as the critical study. The toxic effect detected at this dose is referred to as the critical toxic effect, and the corresponding dose is referred to as the lowest observed adverse effect level (LOAEL). The responses in the critical study obtained at doses below the LOAEL are examined to verify that they constitute NOAELs. The RfD is calculated

---

[2]The remainder of this report will refer only to RfDs; however, the discussion is equally applicable to RfCs.

by dividing the largest NOAEL from the critical study by appropriate uncertainty factors. Table 1 presents uncertainty factors prescribed in the EPA Superfund risk assessment guidance document (U.S. EPA, 1989).

2.3. Overview of the Benchmark Approach

A BMD is defined as a statistical lower confidence limit on the dose producing a predetermined level of adverse change in response compared to the response in untreated animals (the BMR). For example, a BMD could represent a 95 percent statistical lower confidence limit on the dose corresponding to a 1 percent increase in an adverse response over that found in untreated animals. The benchmark level of adverse change in response (the BMR) is 1 percent in this example.

A BMD is calculated by fitting a mathematical dose-response model to data using appropriate statistical procedures. The calculations necessary to determine a BMD are illustrated in Figure 1 for a hypothetical set of dose-response data. The horizontal axis indicates the doses to which the animals were exposed and the vertical axis gives the percent of animals having a particular adverse response. Each solid dot represents the outcome in an experimental dose group. For simplicity, it was assumed that the adverse effect did not occur in untreated animals. The figure depicts a mathematical dose-response model fit to the data and a corresponding curve (also derived from the mathematical model), of statistical lower bounds on doses corresponding to various levels of increased response. The predetermined level of increased response (the BMR) used to define the BMD is shown on the response (vertical) axis. The resulting BMD plotted on the dose (horizontal) axis is determined as the lower bound on dose corresponding to an increased response equal to the BMR. Figure 1 also shows the NOAEL calculated from these data. Although in this particular hypothetical example the BMD is illustrated as being smaller than the NOAEL, a BMD can be either less than or greater than the corresponding NOAEL.

The BMD approach was designed to address the criticisms of the NOAEL approach presented earlier. Those criticisms are largely quantitative or statistical in nature. Thus, the BMD approach is intended to be an approach with statistical properties that are superior to those of the NOAEL approach. The goals of the BMD approach include providing flexibility with respect to the definition of the BMD (i.e., not to be restricted to one of the experimental dose levels) and accounting more appropriately for sample size and dose-response characteristics (Crump, 1984; Dourson et al., 1985; Kimmel and Gaylor, 1988).

It is equally important to be clear about what the BMD approach is not intended to do. Even though mathematical models are used in the approach, the BMD approach is not intended to be used for "low-dose extrapolation," that is, to quantitatively estimate risks at doses far below the range for which increased responses can be directly measured. Since the models proposed for the BMD approach are statistical models that do not incorporate detailed information on the mechanisms through which the toxin causes the particular adverse effect being modeled, their predictions may be seriously in error if used to extrapolate to low doses.

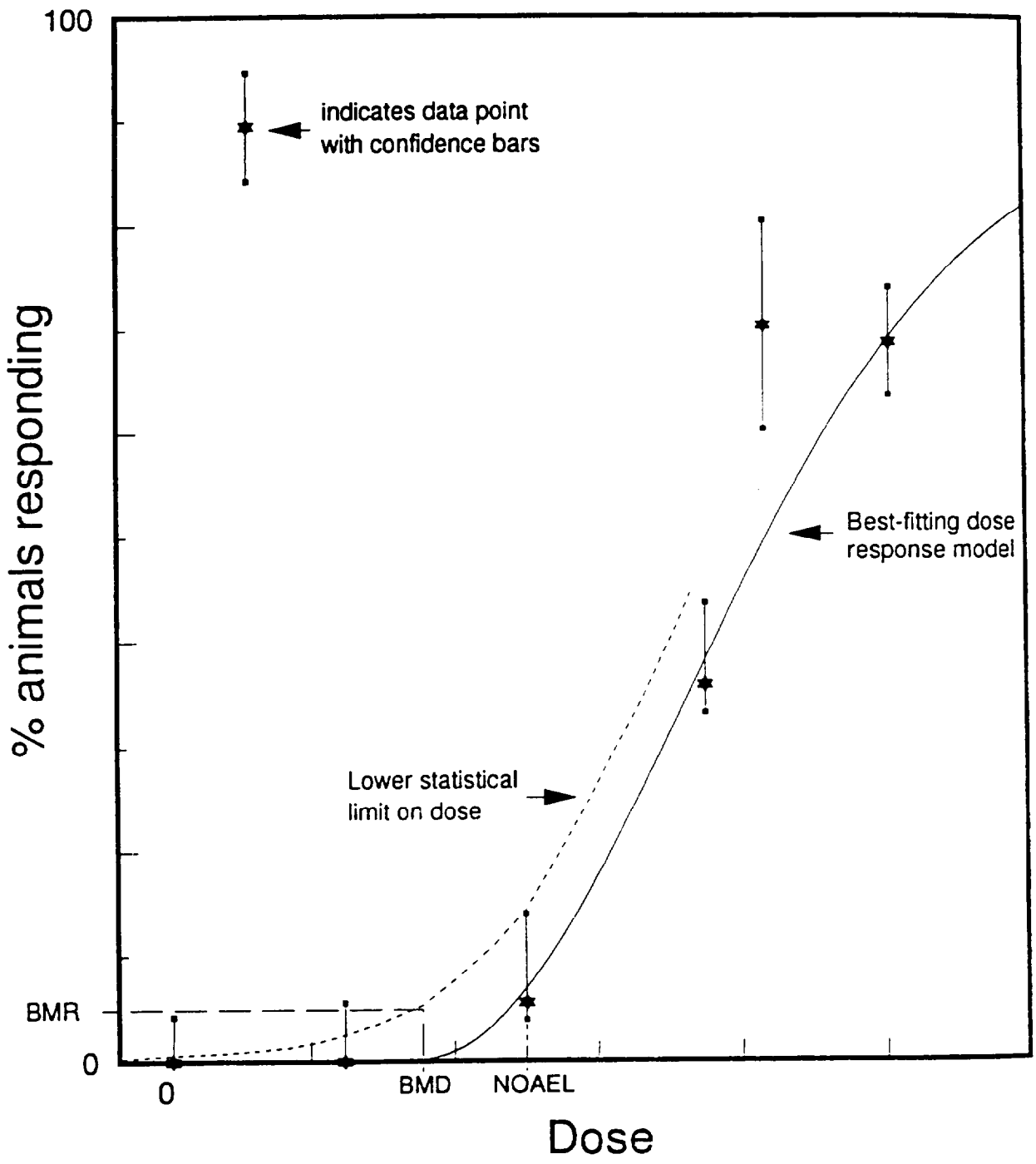
On the other hand, since the calculation of a BMD does not involve extrapolation far beyond the range of the experimental data, it should not be highly dependent upon the dose-response model used. Because of this, there would be little advantage to using detailed

Table 1. Uncertainty Factors

Factor	Value	Description
10H	10	Accounts for variation in the general human population; intended to protect sensitive subpopulations (e.g., elderly, children).
10A	10	Accounts for extrapolation from animals to human; intended to account for interspecies variability between humans and other mammals.
10S	10	Used when a NOAEL is from a subchronic study but a chronic RfD is desired.
10L	10	Used when a LOAEL is used instead of a NOAEL; intended to account for extrapolation from LOAELs to NOAELs.

Source: U.S. EPA, 1989.

Figure 1. Example of Calculation of a BMD



BMR=target response level used to define BMD

mathematical models of underlying biological processes to calculate BMDs, even if such models were available and validated.

3. DETAILED DESCRIPTION OF THE BMD APPROACH

The determination of an RfD using the BMD approach involves three basic steps. First, a response or group of responses from one or more experiments is selected. Second, BMDs are calculated for the selected responses. Third, a single BMD is determined from among those calculated and an RfD is calculated by dividing that BMD by appropriate uncertainty factors. Each of these steps involves a number of decision points that will be discussed in detail in this section.

In the first step, one must decide how to select the experiments and responses for calculating BMDs. In the third step, the values of the uncertainty factors must be chosen. These selections and decisions are required in both the NOAEL and BMD approaches.

Particular attention will be focused here on the second step, which is unique to the BMD approach. This step involves specifying the form in which the data will be recorded for modeling, choosing a dose-response model, selecting the mathematical definition of altered response, stipulating the benchmark level of altered response (the BMR) used to define the BMD, and selecting the procedure for computing statistical confidence limits used to calculate the BMD (including selection of the size of the confidence limit).

Each of the decision points that are required in the BMD approach is listed in Table 2. These decision points and the options available for those decisions are discussed in detail in the following sections. In application of the BMD method, EPA may find it desirable to provide guidance for choosing among these options so that RfDs obtained using the BMD approach are calculated in a consistent manner.

3.1. Selection of Responses to Model

There may be several toxicity studies for a particular substance and each study may contain data for a number of biological effects. In order to calculate a BMD, dose-response models must be applied to one or more effects from one or more studies. Several options for selecting responses for modeling are discussed in the following section.

Certain studies may be eliminated from consideration based on the overall quality of the study, the route of exposure used in the study vis à vis the route of exposure for which an RfD is required, and the range of health effects studied vis à vis those for which the RfD is intended to cover. Such considerations also are used by EPA to focus attention on more relevant studies when calculating NOAELs (U.S. EPA, 1989).

Additionally, specific responses in studies may be eliminated from consideration if there is no convincing evidence of a dose effect for those responses. Such a determination may be based upon the opinions of those who conducted the experiment, possibly supplemented by additional statistical tests.

Table 2. Steps and Decisions Required in the BMD Approach

Step	Decisions
1. Study/Response Selection	• Experiments to include • Responses to model
2. Calculate BMD(s)	• Format of data • Mathematical model(s) • Handling lack of fit • Measure of altered response • BMR definition • Confidence limit calculation
3. Determine RfD	• Specific BMD for RfD calculation • Uncertainty factors

One option would be to apply dose-response models to all of the remaining responses. While this option has the advantage of completeness, it may require a large effort if the data base is sizable. Further, it may be difficult to interpret results from a large number of dose-response analyses.

An option for further limiting the number of responses for modeling is to limit attention to a single critical study, as EPA does in the NOAEL approach (U.S. EPA, 1989). The critical study is generally the one employing the lowest dose (the LOAEL) at which a toxic effect is detected.

In addition, one could choose to model only the effect(s) seen at the LOAEL. This option would minimize the number of responses for which dose-response modeling would be required. However, unlike the calculation of a NOAEL, the calculation of a BMD takes into account the slope of the dose response. Thus, it is possible that an effect seen only at doses above the LOAEL, but having a shallow dose response, could produce a lower BMD than an effect seen at the LOAEL, but having a steeper dose response. This is a potential drawback to modeling only effects seen at the LOAEL.

3.1.1. Example

Sanders et al. (1974) tested the effects of dietary exposure to Aroclor 1254, a PCB mixture, on several biological responses in male albino mice (ICR strain). The researchers examined effects after 2 weeks of exposure on pentobarbital-induced sleeping time; food consumption; serum corticosterone; and weights of the liver, testes, preputial glands, adrenal glands, and vesicular glands. Serum corticosterone levels were elevated for all doses tested (62.5, 250, and 1,000 ppm)[3], pentobarbital-induced sleeping time and food consumption were reduced, and liver weight was increased at 250 and 100 ppm. Adrenal glands were significantly heavier only at 1,000 ppm. Weights of testes, preputials, and vesicular glands were not significantly affected by the PCB ingestions under investigation in this study.

Assume that the study of Sanders et al. (1974) is selected as a study that will be considered for RfD estimation. Since the responses related to weight changes of the testes, preputials, and vesicular glands showed no response to dose, these responses might be ignored for the purposes of BMD estimation. If one chose to model only responses seen at the LOAEL (62.5 ppm in this case), serum corticosterone level would be the only response parameter modeled. Otherwise, serum corticosterone, liver weight, adrenal gland weight, pentobarbital-induced sleeping time, and food consumption could be modeled because, depending on the slope of the dose-response curves, any one of these responses could yield the smallest BMD for this study.

When considering this study in the context of other studies selected for RfD estimation, some or all of the responses observed by Sanders et al. (1974) might not be modeled if the 62.5 ppm dose (suitably transformed to yield consistent units across all studies) was not the LOAEL among all the studies (i.e., Sanders et al. [1974] is not the critical study). However, even in the

[3]Although a higher dose of 4,000 ppm was tested, all mice exposed at that level died within 7 days of initial exposure.

more general context of all relevant PCB studies, one of the responses from Sanders et al. (1974) could yield the smallest BMD, again depending on the dose-response slopes and the doses used in the other studies.

The choice among the responses in this study also might be limited by consideration of the relevance of the responses to the RfD that is to be estimated.

3.2. Format of Data to Be Used for Modeling

The type of dose-response model that is suitable for modeling a set of data depends in part on the format in which the data are recorded. This format may also affect the numerical value of the BMD obtained.

Non-cancer health effects can be recorded in either categorical or continuous formats. In a categorical format, possible responses are divided into two or more groups and the numbers of responses in each group are recorded. For example, organ degeneration may be recorded as absent, mild, moderate, or severe. The most commonly used format for categorization of data is the quantal format in which only the presence or absence of the response in an experimental subject is noted. At the other extreme, a response may be capable of assuming a continuum of values and be recorded in a continuous format. Organ weights and serum enzyme levels are examples of responses that are often recorded in a continuous format.[4]

The format used for expressing a response may be determined largely by what is customary or appropriate for a particular type of response. For example, cancer responses and particular types of developmental effects are generally recorded in a quantal form simply as present or absent without more detailed categorization.

Additionally, unless there is access to the raw data from a study, the format for expressing a response will be limited by the format in which the data are summarized. Clearly, data cannot be categorized more finely than in the data summary available. When the raw data for a response in question are available in a continuous format, either they can be used directly in a continuous format in the dose-response models, or they can be converted into a categorical format by dividing the range of the responses into subintervals and recording the number of subjects with responses in each subinterval. For certain continuous responses, a particular interval in the range may be considered to represent the "normal range" for this response. Normal ranges can be used to define corresponding quantal responses in a very natural fashion by considering a subject to be affected if its response is outside of the normal range.

---

[4]An additional possibility is for a response to be reported in a format that is a combination of continuous and categorical. Consider, for example, the measurement of a serum enzyme level by an analytical method that has a detection limit of x micrograms per liter. Subjects with a response higher than x would have their response recorded continuously; whereas for subjects with a response less than that, the detection limit only would have their response categorized
as $<x$.

It may be preferable in some cases to recode continuous data in a quantal form because a quantal format relates more directly to adverse response, which is the basis for RfD determination. Consider, for example, the response of liver weight as a fraction of total body weight. Liver weight as a fraction of total body weight is not adverse per se; however, it may represent an adverse response when it reaches a certain level. If this level was specified, then animals with liver weight to total weight ratios above that level could be considered to be adversely affected. This would define a quantal response to which a quantal BMD approach could be applied. As another example, since a body weight reduction of $\geq$10 percent has been defined as an adverse effect (OSTP, 1985), body weight changes could be treated quantally using the 10 percent cutpoint to define the presence of an adverse response.

A disadvantage of recoding continuous data into a categorical form is that information on the magnitude of the response is lost. An advantage is that the categories may be defined to correspond to normal and abnormal ranges and therefore permit the response to be more easily interpreted in terms of an adverse effect. Another possible advantage is that, since generally some of the responses of interest must be categorical, comparisons among responses may be facilitated if they are all categorical, and particularly if they are all quantal. On the other hand, the data needed to define a categorical response may not be available in a published report.

3.2.1. Example

Johnson et al. (1986) examined the effect in rats of chronic acrylamide exposure on degeneration of tibial nerves. The degree of degeneration (from very slight to severe) was recorded for each rat. Data of this form are categorical but not quantal. Because degeneration of the type observed has been observed in aging rats (Johnson et al., 1986) and because very slight and slight degeneration was observed at roughly the same rate in all dose groups, adverse effect was defined to be moderate or severe degeneration. This definition also defines a quantal response, with degeneration that was slight or very slight counting as no response and moderate and severe degeneration counting as a response; the numbers of male rats with moderate or severe degeneration are displayed in Table 3.

3.3. Mathematical Models for Defining a BMD

A mathematical dose-response model must be selected to use in estimating the BMD. Different types of models are required for categorical and continuous data and these different types of models have different data requirements.

In the case of categorical data, the information generally required for application of dose-response models includes the experimental doses, the total number of animals in each dose group, and the number of these whose responses are in each of the categories. We will generally be interested in the special case of quantal responses (i.e., two categories), and only models for this special case will be discussed. The information required for application of dose-response models to quantal data is the experimental doses, the total number of animals in each dose group, and the number in each group with the response of interest.

In the case of continuous data, for application of a number of dose-response models (specifically, those that assume that responses at each dose level are normally distributed)

Table 3. Acrylamide-Induced Tibial Nerve Degeneration in Rats

Data	Dose (mg/kg/day)	Number Affected	Number Tested
	0	9	60
	0.01	6	60
	0.1	12	60
	0.5 (NOAEL)	13	60
	2.0	16	60

Modeling Results

Model	Goodness-of-Fit p-Value	BMD (mg/kg/day) (5% extra risk)
QQR	0.34	0.83
QW	0.48	0.31

QQR = quantal quadratic regression
QW = quantal Weibull

Source: Johnson et al., 1986.

experimental doses, the number of animals in each dose group, the mean response in each group, and the sample variance of the response in each group must be known.

Various models that have been proposed for quantal and continuous data (Crump, 1984; Gaylor, 1989; Gaylor and Slikker, 1990) are listed in Table 4. Each of these models involves three or more parameters that are estimated by fitting the model to experimental data. This fitting is usually accomplished by a statistical procedure known as maximum likelihood (see Appendix A). This procedure provides estimates of the parameters, and from these estimates, the probability of response (for quantal data) or the mean response (for continuous data) can be estimated for each dose level. The maximum likelihood procedure also can be used to compute a lower statistical confidence limit for the dose corresponding to the BMR. This lower confidence limit is defined as the BMD.

The models shown here, as well as many other possible models, relate the response to the level of dose, d. The response variable is denoted in Table 4 either by P(d), the probability of response, for a disease outcome that is either present or absent (quantal), or by m(d), the mean value of a continuously measured parameter of health or well-being. In all of the equations shown, d_0 is a threshold dose level, a dose level below which the response variable is unaffected (i.e., at doses less than or equal to the threshold, the response variable value remains at c, the value of that variable in the absence of dosing). For the quantal models, the probability of response is assumed to increase as dose level increases. For the continuous models, mean response can either increase or decrease as a function of dose level.

The models listed in Table 4 are statistical models for describing quantitatively the pattern of biological effects within the range of the observed data. The models are proposed for application to a wide range of effects with diverse underlying biological mechanisms, and they do not model detailed information regarding underlying mechanisms.

Several criteria that may be considered when selecting a dose-response model to calculate a BMD are discussed below.

Ability to Describe the Observed Dose Response. Since the goal of the BMD approach is the estimation of a lower bound on dose for some level of risk not far below the observed range, the model should give adequate predictions of the observed experimental responses. Goodness-of-fit tests (see Appendix A) can be applied to determine if a model adequately describes the dose-response data.

Each of the models presented in Table 4 is capable of describing a range of dose-response patterns. Figures 2 through 6 give an indication of the dose-response curve shapes that may be obtained with some of these models. The QPR and LN models will provide dose-response shapes similar to that shown for the QW model (Figure 4). Similarly, the CPR model will provide a range of patterns similar to that shown for the CP model (Figure 6). Although Figures 5 and 6 depict the CQR and CP models applied to a response that decreases as the dose increases, these and all of the models for continuous data also can be applied to responses that increase with increasing dose.

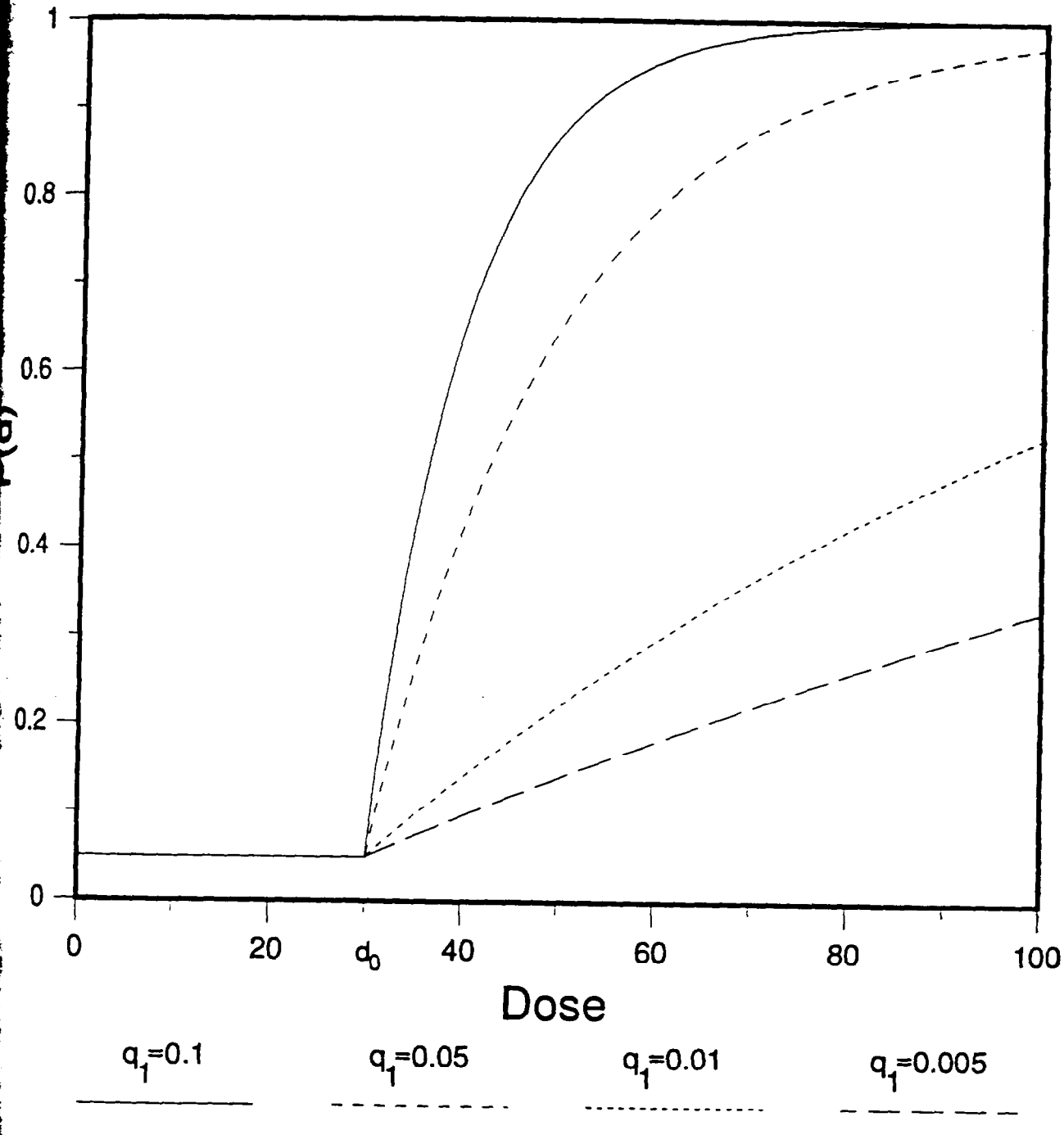
Table 4. Dose-Response Models Proposed for Estimating BMDs

Model	Formula
Quantal Data	
Quantal linear regression (QLR)	$P(d) = c + (1-c)\{1-\exp[-q_1(d-d_0)]\}$
Quantal quadratic regression (QQR)	$P(d) = c + (1-c)\{1-\exp[-q_1(d-d_0)^2]\}$
Quantal polynomial regression (QPR)	$P(d) = c + (1-c)\{1-\exp[-q_1d-...-q_kd^k]\}$
Quantal Weibull (QW)	$P(d) = c + (1-c)\{1-\exp[-q_1d^k]\}$
Log-normal (LN)	$P(d) = c + (1-c)N(a+b\ \log d)$
Continuous Data	
Continuous linear regression (CLR)	$m(d) = c + q_1(d-d_0)$
Continuous quadratic regression (CQR)	$m(d) = c + q_1(d-d_0)^2$
Continuous linear-quadratic regression (CLQR)	$m(d) = c + q_1d+q_2d^2$
Continuous polynomial regression (CPR)	$m(d) = c + q_1d+...+q_kd^k$
Continuous power (CP)	$m(d) = c + q_1(d)^k$

Note: P(d) is the probability of a response at the dose, d; m(d) is the mean response at the dose, d. In all models, c, $q_1,...q_k$, and d are parameters estimated from data. For the quantal models, $0 \le c \le 1$ and $q_i \ge 0$. For the CPR model proposed by Crump (1984), all the q_i have the same sign. In the CLQR model discussed by Gaylor and Slikker (1990), q_1 and q_2 were not constrained to have the same sign. For all models, $d_0 \ge 0$, $k \ge 1$. N(x) denotes the normal cumulative distribution function.

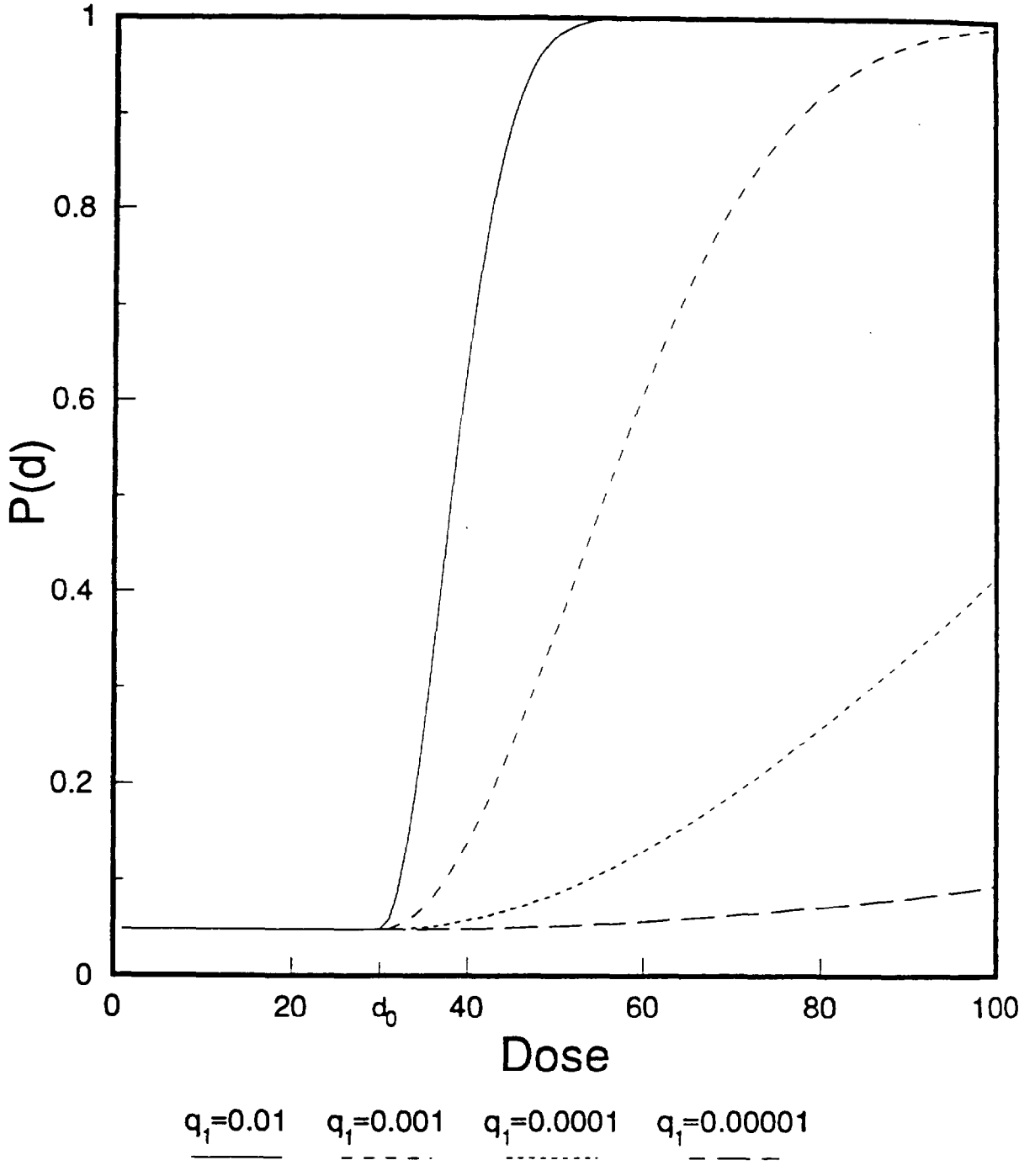
Source: Crump, 1984; Gaylor, 1989; Gaylor and Slikker, 1990.

Figure 2. **Examples of Quantal Linear Regression (QLR) Curves:**
$P(d)=c+(1-c)\{1-\exp[-q_1(d-d_0)]\}$



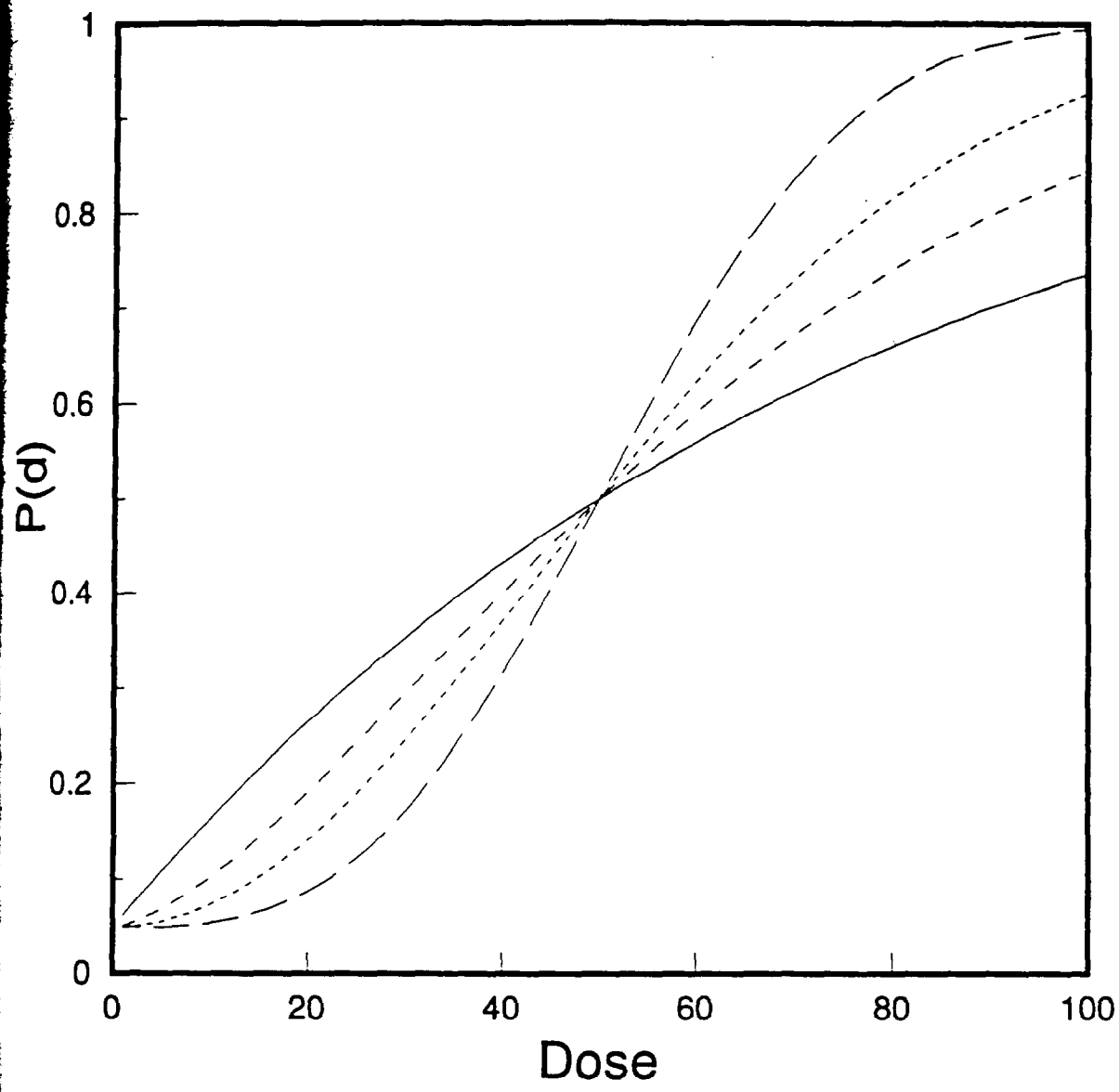
All curves show maximum likelihood predictions of response, not confidence limits, for various choices of parameters. For all curves, c=0.05, d_0=30. The parameter q_1 is the dose coefficient in this model; larger values of q_1 give steeper dose response.

Figure 3. **Examples of Quantal Quadratic Regression (QQR) Curves:**
$P(d)=c+(1-c)\{1-\exp[-q_1(d-d_0)^2]\}$



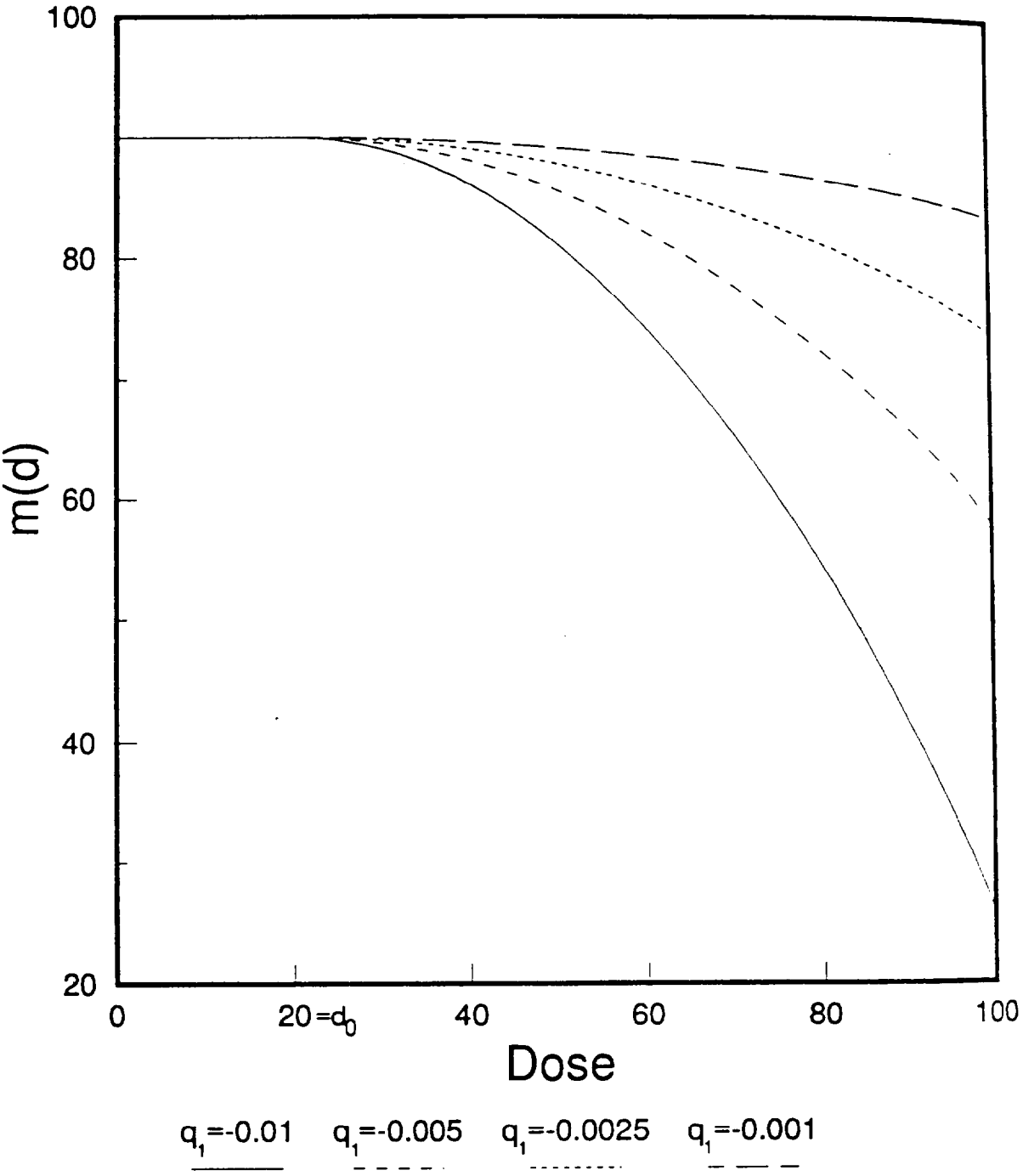
All curves show maximum likelihood predictions of response, not confidence limits, for various choices of parameters. For all curves, c=0.05, d_0=30. The parameter q_1 is the dose coefficient in this model; larger values of q_1 give steeper dose response.

Figure 4. **Examples of Quantal Weibull (QW) Curves:**
$P(d)=c+(1-c)\{1-\exp[-q_1(d)^k]\}$



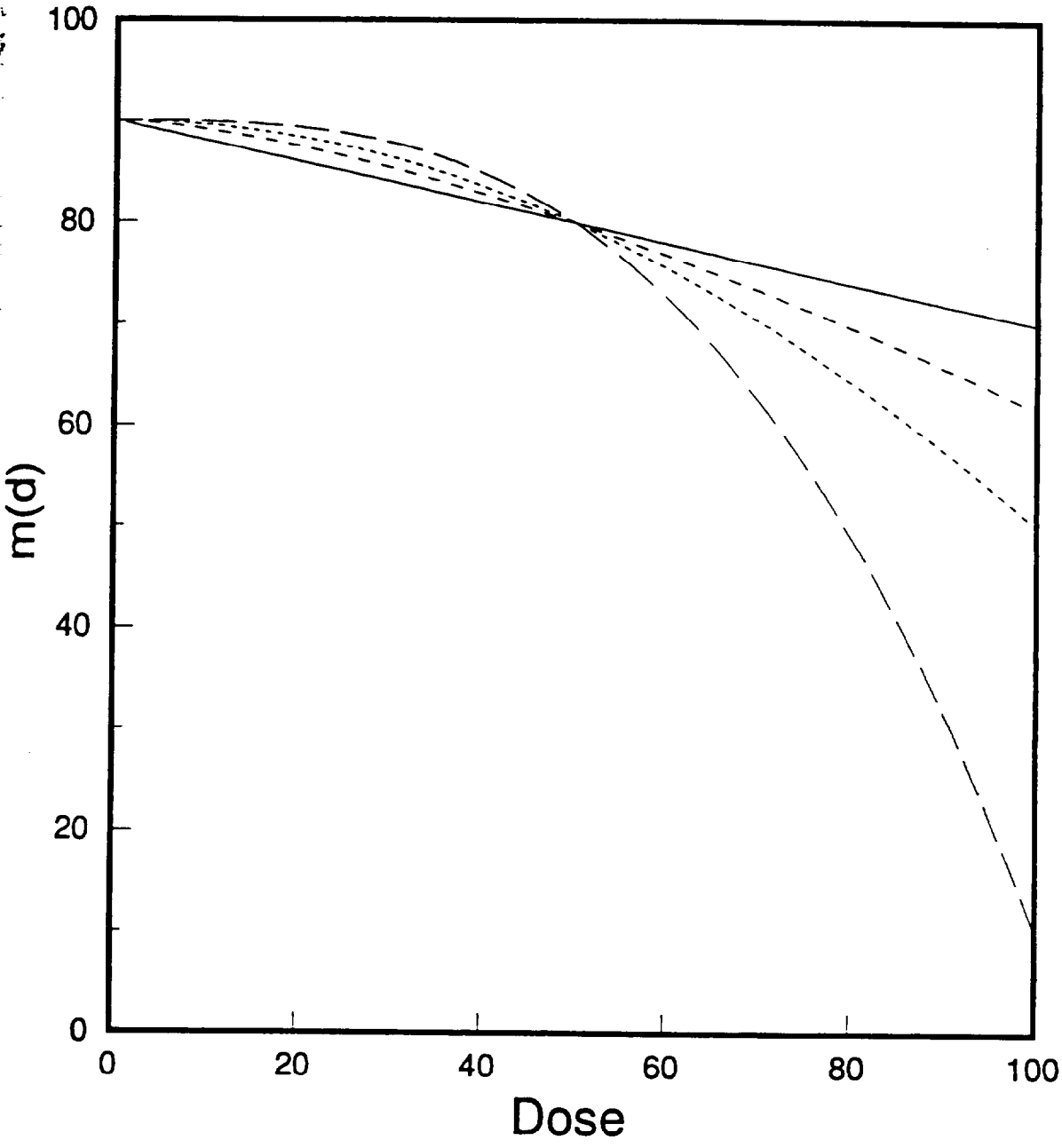
All curves show maximum likelihood predictions of response, not confidence limits, for various choices of parameters. For all curves, c=0.05. The parameter q_1 is the dose coefficient in this model; larger values of q_1 give steeper dose response. The parameter k is the power on dose; larger values of k give more curvature.

Figure 5. **Examples of Continuous Quadratic Regression (CQR) Curves: $m(d)=c+q_1(d-d_0)^2$**



All curves show maximum likelihood predictions of response, not confidence limits, for various choices of parameters. For all curves, $c=90$, $d_0=20$. The parameter q_1 is the dose coefficient in this model; larger values of q_1 give steeper dose response.

Figure 6. Examples of Continuous Power (CP) Curves: $m(d)=c+q_1(d)^k$



All curves show maximum likelihood predictions of response, not confidence limits, for various choices of parameters. For all curves, c=90. The parameter q_1 is the dose coefficient in this model; larger values of q_1 give steeper dose response. The parameter k is the power on dose; larger values of k give more curvature.

It is often the case that several models will adequately describe the data under consideration. When that is true, other considerations must be used to decide on the model to use for BMD calculation.

Statistical Assumptions. An important consideration in selecting a model is the reasonableness of the statistical assumptions underlying a model and the procedures used to fit it to the data. In most instances, it may be reasonable to assume that quantal results arise from binomial variation about a dose-dependent expected number of responders. This means that each subject is assumed to respond independently of all other subjects, and that all animals in a given dose group have an equal probability of responding. These assumptions are generally made when applying the models for quantal data listed in Table 4. Similarly, a continuous endpoint is generally assumed to display variation in accordance with dose-dependent normal distributions. In other words, each subject is assumed to respond independently of all other subjects, and the responses of animals in a particular dose group are distributed according to a normal probability distribution. The methods proposed by Crump (1984) for fitting the continuous models listed in Table 4 assume this type of normal variation.

There are situations, however, when the binomial or normal assumptions may not be appropriate. In those cases, one should consider alternative models that are based on more appropriate assumptions.

One example of this is in studies of developmental toxicity where responses within and across litters are observed. In such experiments, the response in one fetus may not be independent of the response in other fetuses in the same litter. Consequently, the assumption of independence inherent in models that assume binomial variation is not strictly valid, although this assumption may still provide reasonable results in specific cases.

Alternative models that assume more general forms of variation for quantal responses from developmental toxicity experiments have been developed (Rai and Van Ryzin, 1985; Kupper et al., 1986; Kodell et al., 1991). Such models should be considered when applying the BMD approach to responses observed in individual fetuses.

As a different example, it may be necessary to transform continuous data in some cases so that they better satisfy the assumptions of a normal distribution. A log-transform is often used for this purpose. A decision regarding whether and how to transform continuous data is another possible choice required when modeling continuous data. Kendall (1951) presents statistical tests that can be used to determine if data are consistent with a normal assumption, and Steel and Torrie (1980) discuss data transformations that can be considered to make the data more normal. Generally, one will need to have access to the raw data from an experiment in order to make a data transformation.

Biological Considerations. Even though the models in Table 4 are descriptive and do not incorporate detailed information on biological mechanisms, certain general biological considerations may be used to help select the dose-response models to be used for BMD calculation.

One example could be in selection of a threshold versus a non-threshold model. The quantal models QLR and QQR involve a threshold dose, d_0. Doses below this threshold[5] are assumed not to affect the probability of a response. On the other hand, the quantal models QPR, QW, and LN do not involve a threshold dose, and consequently with these models any dose, no matter how small, is assumed to increase the probability of an adverse response. One possible input for selection of a model is to apply threshold models to responses that are thought likely to have thresholds, and non-threshold models for responses for which thresholds are considered less likely, based on consideration of biological mechanisms.[6]

However, since a BMD is a dose corresponding to a finite (non-zero) increment in response (the BMR), even if a threshold exists, the model predictions are only used for doses that are above the threshold. In applying both threshold and non-threshold models to several data sets, Crump (1984) did not find large differences between BMDs calculated from models involving thresholds and those not involving thresholds. Indeed, one goal in the selection of a BMR is for the resulting BMD not to be highly dependent upon the underlying model. If this goal is accomplished, then it should make little practical difference whether the model used includes a threshold.

Biological considerations also might be used to select models based on the biological plausibility of the dose-response curve shape. Consider, for example, the difference between the QLR and QQR models (Table 4) at doses near the threshold dose, d_0. While the QLR model has a sharp transition from the background response rate to the dose-dependent rate at the threshold, the transition for the QQR model is smoother, without the apparent abrupt change (compare Figures 2 and 3). In some circumstances, a smooth (continuous) change of slope may be deemed more reasonable for the response under consideration and the QQR model favored over the QLR model. In this case as well, however, if the BMR is selected appropriately (i.e., large enough so that lower bounds on dose for that level of risk are not overly dependent on the choice of model), it should make little practical difference which of these models is selected.

Use of Multiple Models. It may be difficult to limit the calculations to a single model based on the criteria discussed above. Consequently, it may be desirable to apply several models. More than one of those models may fit the observed responses equally well. The decisions required in that case are discussed in Section 3.8.

---

[5]A distinction is sometimes made between a threshold for an individual and a threshold for a population. Even if thresholds exist, if individual thresholds differ according to some probability distribution, then a single threshold may not apply to a large population (Crump et al., 1976). In the models listed in Table 2 that incorporate a threshold dose, d_0, that single threshold is assumed to apply to all individuals.

[6]The existence or non-existence of a threshold for an effect can never be known with certainty based strictly on experimental results for that effect. If no responses are found at a given dose, it is always possible that another experiment employing larger numbers of animals could detect a response. Conversely, if responses are detected at a given dose, it is always possible that a threshold might exist at some lower dose.

3.3.1. Example

Tibial nerve degeneration induced by acrylamide was observed by Johnson et al. (1986) in male rats. The responses were quantalized as discussed in Section 3.2.1 and shown in Table 3. Table 3 also summarizes the results of fitting two quantal models (the quantal quadratic regression, QQR, and quantal Weibull, QW, models from Table 4). Figure 7 shows the rates of moderate and severe degeneration, the best fitting QQR model, and the best fitting QW model.

Both models fit the data satisfactorily; chi-squared goodness-of-fit tests yielded nonsignificant p-values (see Appendix A). The QW model provided a slightly better fit to the data.

If the QQR and QW models were the models that were being considered for BMD estimation for quantal responses, then a choice of one or the other would be necessary, unless some procedures for dealing with multiple BMD estimates were adopted (see Section 3.8). The fits of the models to the data are both adequate and the statistical assumptions underlying the two models are identical; these considerations do not suggest acceptance of one model over the other. The QW model does not allow a threshold, whereas the QQR model does. If it is suspected that tibial nerve degeneration does not have a threshold, then one might prefer to use the QW model. If the opposite is the case, then the QQR model might be preferred. The observance of high rates of very slight and slight degeneration, and non-zero rates of moderate and severe degeneration, in the control group of Johnson et al. (1986), in addition to other biological considerations, may suggest whether a threshold assumption is reasonable and consequently may help determine the choice of models.

3.3.2. Additional Research

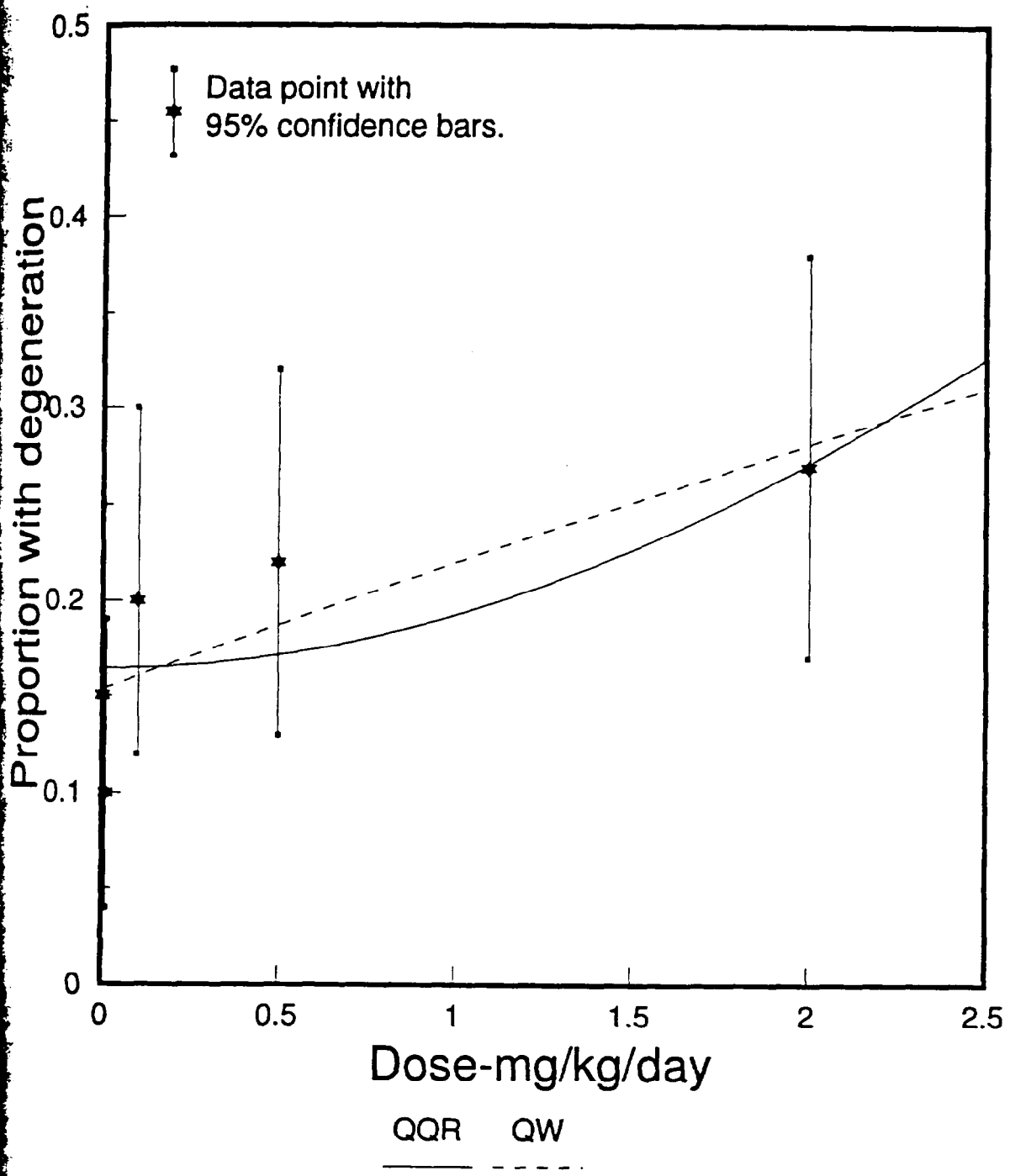
The types of dose-response models reviewed herein are not appropriate for all forms of toxicological data. Several types of data may require other types of models. As described above, studies of developmental toxicants observe response in fetuses. Different fetuses from the same litter do not respond independently to developmental toxicants, and models that account for such "litter effects" may be needed.

Endpoints exhibiting different severities constitute another type of data that may require special modeling approaches. Sometimes, in addition to knowing whether an animal was affected, the level of effect may be categorized (e.g., mild, moderate, severe). While such categorization can be ignored, it could be useful to have models available for calculating BMDs that can take advantage of the additional information.

In some studies several different durations of exposure can be used. Correspondingly, there may be a need to set different RfDs for different durations of human exposure. To accomplish this, models that incorporate duration of exposure as well as the dose level may be required.

Some models exist that are applicable to each of these situations (see Rai and Van Ryzin, 1985; Kodell et al., 1991; Clement International Corporation, 1990a, b). However, the applicability of such models to calculating BMDs needs to be studied. Additional models and the software needed to implement them may need to be developed as well.

Figure 7. Moderate to Severe Nerve Degeneration in Rats Following Acrylamide Exposure



Source: Johnson et al., 1986.

3.4. Adjusting for Lack of Fit

None of the models listed in Table 4 will provide a reasonable fit to certain data sets. Frequently, this is due to reduced responses at higher doses that are inconsistent with the dose-response trend seen at lower doses. One likely reason for this is interference at higher doses by competing mechanisms of toxicity. Whenever a lack of fit occurs, one should be sure that all affected animals are being taken into account. For example, in some experiments, if a high incidence of response is seen at lower doses, the experimenter may not look for the effect at higher doses. As another example, suppose a BMD is being calculated based on the response, "mild atrophy." If mild atrophy progresses to "moderate atrophy" and subsequently to "severe atrophy," then animals with these more severe forms should be considered to be affected as well. In general, if a BMD is being calculated based on a toxic response that can progress to more severe forms (possibly known by different names than the original response), animals with more severe forms of the response should be also considered to be affected.

In other cases, a particular response may be reduced at higher doses due to interference by other responses that are not a progressive form of the response of interest. One such example is when a dose-related toxic response that occurs primarily in aged animals is not expressed because of premature deaths due to other toxic effects. A more subtle example would be if moderate doses caused a particular organ to be enlarged, but still higher doses caused the same organ to atrophy through an independent mechanism. In these cases, it would probably not be appropriate to combine these separate toxic responses into a common response.

Whenever the responses at the higher doses are reduced, so that none of the models listed in Table 4 fit, one option would be to look for a more flexible model that can adequately describe the dose response. A seeming advantage to this approach is that one may be able to incorporate all of the data into the analysis. A danger in this approach is that the attempt to fit the high-dose data will skew the dose response at the lower doses that are of more direct interest.

A simpler and perhaps better advised approach is to omit the data at the highest dose when none of the models provide an adequate fit, and refit the models to the remaining data. This process can be continued and an adequate fit will eventually be obtained.[7] This approach is used by EPA in risk assessments for cancer based on the linearized multistage model (Anderson et al., 1983). The rationale for eliminating data at the highest dose as opposed to lower doses is that the data at the highest dose should be the least informative of responses in the lower dose region of interest.

A plateau in the responses at the higher doses can be caused by saturation of metabolic or delivery systems for the ultimate toxic substance. Such an effect can also cause dose-response models not to fit the data adequately. It may be possible to overcome this problem by estimating the delivered dose to the site of action, and then applying these doses in the dose-response models rather than an external measure of exposure (Andersen et al., 1987). In this

[7]The only exception to this is if there is a statistically significant deficit in response at the lowest dose below that seen in control animals; however, it is questionable whether data such as these should be used to determine an RfD.

approach, pharmacokinetic data on animals are used to estimate a measure of internal dose to the target tissue that results from the experimental dosing regimen. The BMD method is applied to these internal measures of dose to estimate an RfD for internal dose. Human pharmacokinetic information is then used to estimate the external dose that would result in an internal dose equal to the internal RfD; this external dose would be defined as the human RfD.

3.4.1. Examples

Ethylene glycol monopropyl ether (EGPE) was examined for toxic effects in rats when administered for 6 weeks via gavage (Katz et al., 1984). As part of the study, the spleen was examined at the end of the 6-week exposure period. In several of the rats, the spleen appeared dark and enlarged, presumably as a result of the exposure. Upon histopathological examination, the spleens of affected rats were found to have congestion or extramedullary hematopoiesis. The rates of response for the hematopoiesis are displayed in Table 5. Note that no animals in the high-dose group had that lesion. This may be due to competing manifestations of toxicity or other unexplained reasons.[8]

The NOAEL for this study was determined by applying a statistical technique referred to as the no statistical significance of trend (NOSTASOT) approach (Tukey et al., 1985). (The NOSTASOT approach is described in some detail in Appendix A).

The NOSTASOT procedure applied to all of the dose groups indicated that there was no significant trend for larger doses to yield larger proportions of responders (the Mantel-Haenszel trend test p-value was about 0.56). However, the NOSTASOT procedure applied to the data both without the highest dose group and without the highest two dose groups detected significant trends. Moreover, the pairwise comparison of the 7.5 mmole/kg dose group and the control group indicated a significantly increased rate of response at 7.50 mmole/kg ($p = 0.04$ by Fisher's exact test). The pairwise comparison of the 3.75 mmole/kg dose group and controls was not significant ($p = 0.11$ by Fisher's exact test).

The finding of no significant trend when the NOSTASOT procedure was applied to all of the data might be interpreted to mean that 15 mmole/kg/day is the NOAEL. However, because of the significant trend observed over doses below 15 mmole/kg/day, and because of the other effects observed in the spleen at 15 mmole/kg/day, it appears more reasonable to select 1.88 mmole/kg/day as the NOAEL.

The application of the BMD approach was also interesting in this case. Neither the QQR model nor the QW model could fit the dose-response data when all dose groups were included (p-values less than 0.02). However, dropping the highest dose group (see Section 3.4) resulted in acceptable fits for both models (Table 5). Figure 8 shows the results of fitting the models to the data, ignoring the highest dose group.

---

[8]Four of the high-dose rats had enlarged and darkened spleens; six high-dose rats had congestion in the spleen. These other endpoints might be used in lieu of extramedullary hematopoiesis for determining an RfD for EGPE, but for the sake of illustration the hematopoiesis response is discussed here.

Table 5. EGPE-Induced Extramedullary Hematopoiesis in the Spleen of Rats

	Dose (mmole/kg/day)	Number Affected	Number Tested
Data	0	0	10
	1.88 (NOAEL)	0	10
	3.75	3	10
	7.50	4	10
	15.0	0	10

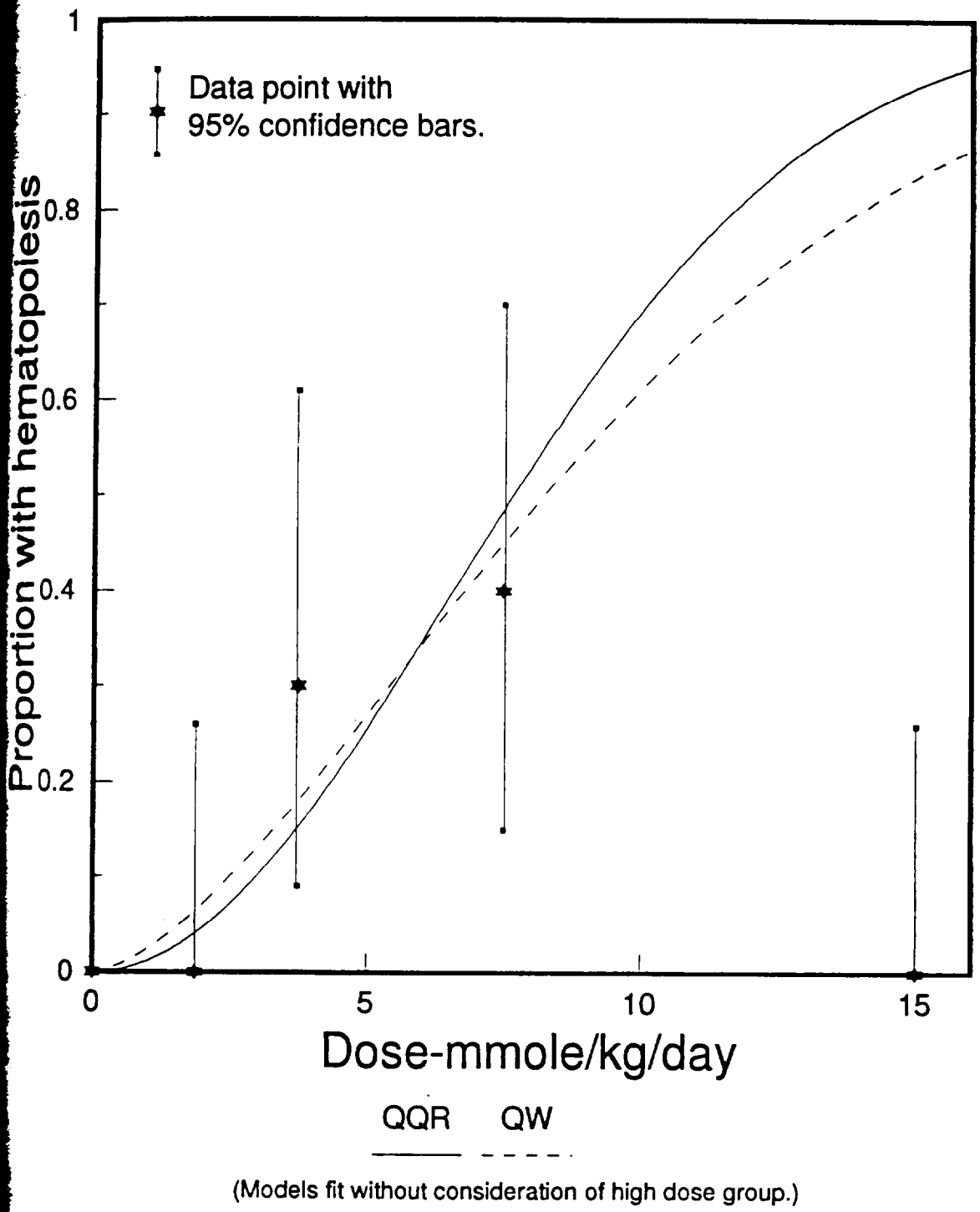
Modeling Results[a]

		BMD (mmole/kg/day) for Extra Risk of:		
Model	Goodness-of-Fit p-Value	10%	5%	1%
QQR	0.30	2.24	1.56	0.69
QW	0.18	0.99	0.48	0.094

[a] The results for the models are those fit to all dose groups except for the highest dose group. Neither model adequately fit data from all dose groups.

Source: Katz et al., 1984.

Figure 8. Extramedullary Hematopoiesis of the Spleen in Rats Following EGPE Exposure



(Models fit without consideration of high dose group.)

Source: Katz et al., 1984.

Table 5 shows the BMD estimates corresponding to three levels of extra risk for the QQR and QW models.

Another example of lack of fit that is not so directly accommodated is provided by a study of glycol ether-induced reproductive toxicity. Miller et al. (1981) examined the effects of 9-day inhalation exposures (6 hours per day) to ethylene glycol monomethyl ether (EGME) on testicular toxicity in rats and mice. Toxicity was determined by measuring testes weights (Table 6). In both rats and mice, testes weights were significantly decreased following exposure to 1,000 ppm. The NOAEL was 300 ppm for both rats and mice.

The best-fitting CQR and CP models are shown in Figures 9 and 10. Although both models fit the rat data adequately, neither model could adequately describe the mouse data (Table 6). The lack of fit to the mouse data is due primarily to the 100 ppm dose group, for which the testes weights were larger (on average), than the controls, and to the small amount of variation in the observed results.

The case of the mouse data illustrates one of the difficulties that can arise in the application of the BMD approach. The lack of fit in this case was not due to reduced response at the highest dose, but rather a reduced response at the low dose. Therefore, dropping dose groups (as discussed in this section) will not lead to an adequate fit.[9]

It is not likely that alternative models will provide better fits to the mouse data, as long as such models postulate a monotone dose response. Models with monotone dose will not be able to predict the increased testes weights in the 100 ppm group. Biological and toxicological considerations may dictate that a non-monotone response pattern is feasible in this instance, in which case one may conclude that doses of 100 ppm or less to male mice do not result in testicular weight loss. Alternatively, it may be determined that the observed variation among the responses underestimates the true variability associated with the testicular response, in which case the predictions of the CQR and CP models may be adequate for the application of the BMD approach.

---

[9]The small standard deviations reported for all the dose group responses entail small estimates of "pure error" used for comparison with the error between model predictions and observations. An F test is performed, where the numerator represents the error for lack of fit and the denominator represents the pure error or the variability of the observed weights around the group-specific means. When the estimate of pure error is small (i.e., when standard deviations are small), deviations of the model predictions from the observations may be significant, even when they appear to be in fairly close agreement.

In some instances values may be erroneously recorded as standard deviations, when in fact they represent standard errors of the means. Whenever this occurs, there is more variability in the observations than suggested by the reported standard deviations, and the models may provide a satisfactory fit. The best insurance against such an error is to have available the results in individual animals. In this case, if the values reported by Miller et al. (1981) are actually standard errors of the means, the CQR and CP models would adequately fit the mouse data.

Table 6. EGME-Induced Testicular Toxicity in Rats and Mice

Data

	Rats[a]		Mice[a]	
Dose (ppm)	Average Weight	SD	Average Weight	SD
0	2.82	0.10	0.21	0.01
100	2.88	0.05	0.23	0.01
300 (NOAEL)	2.70	0.20	0.20	0.02
1000	1.50	0.10	0.10	0.01

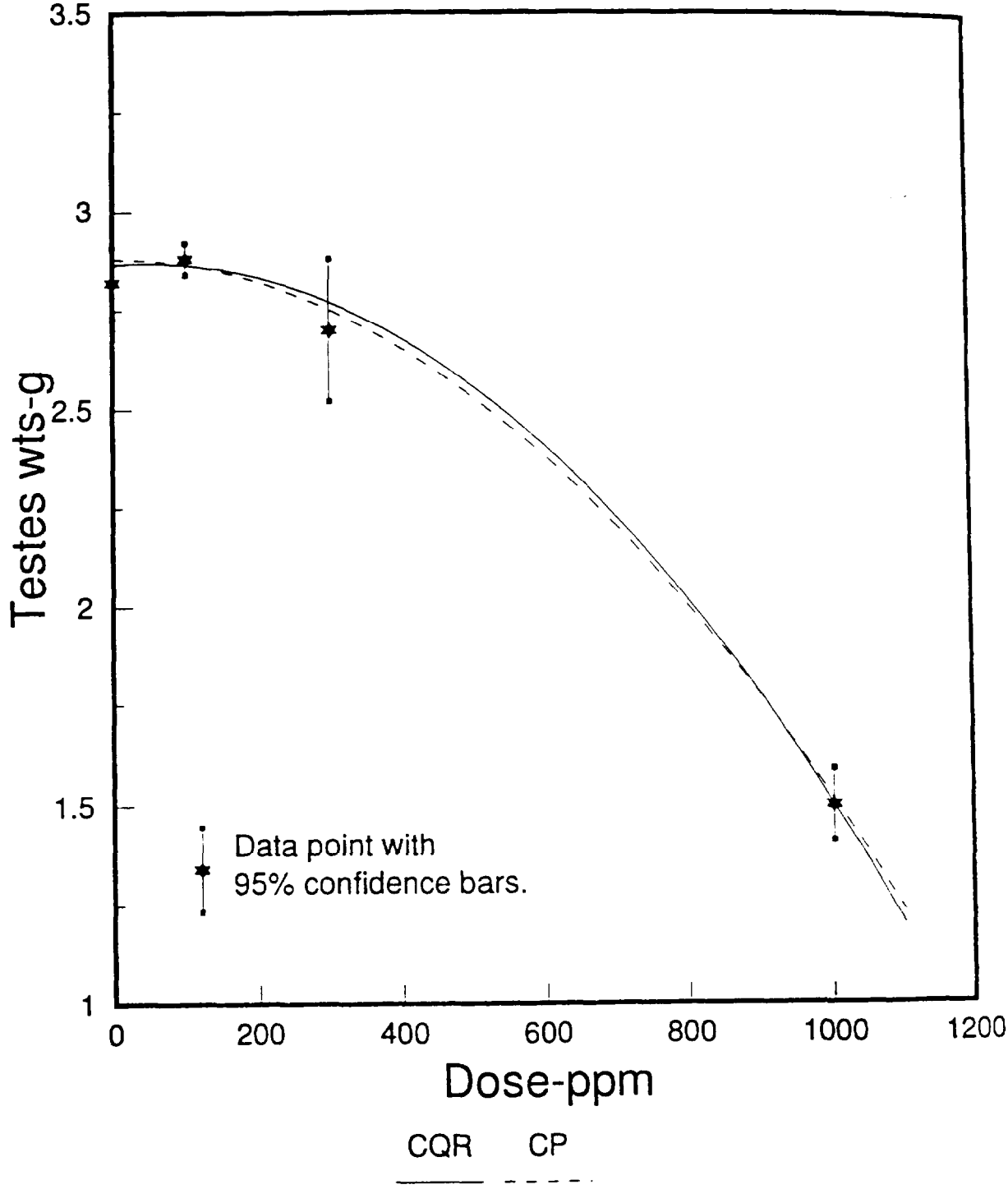
Modeling Results

	Rats		Mice	
Model	Goodness-of-Fit p-Value	BMD (ppm)	Goodness-of-Fit p-Value	BMD (ppm)
CQR	0.13	315	<0.01	---
CP	0.17	184	<0.01	---

[a] Five animals in each dose group. Reported are average testes weights (in grams) and standard deviations (SD) for testes weights in each dose group.

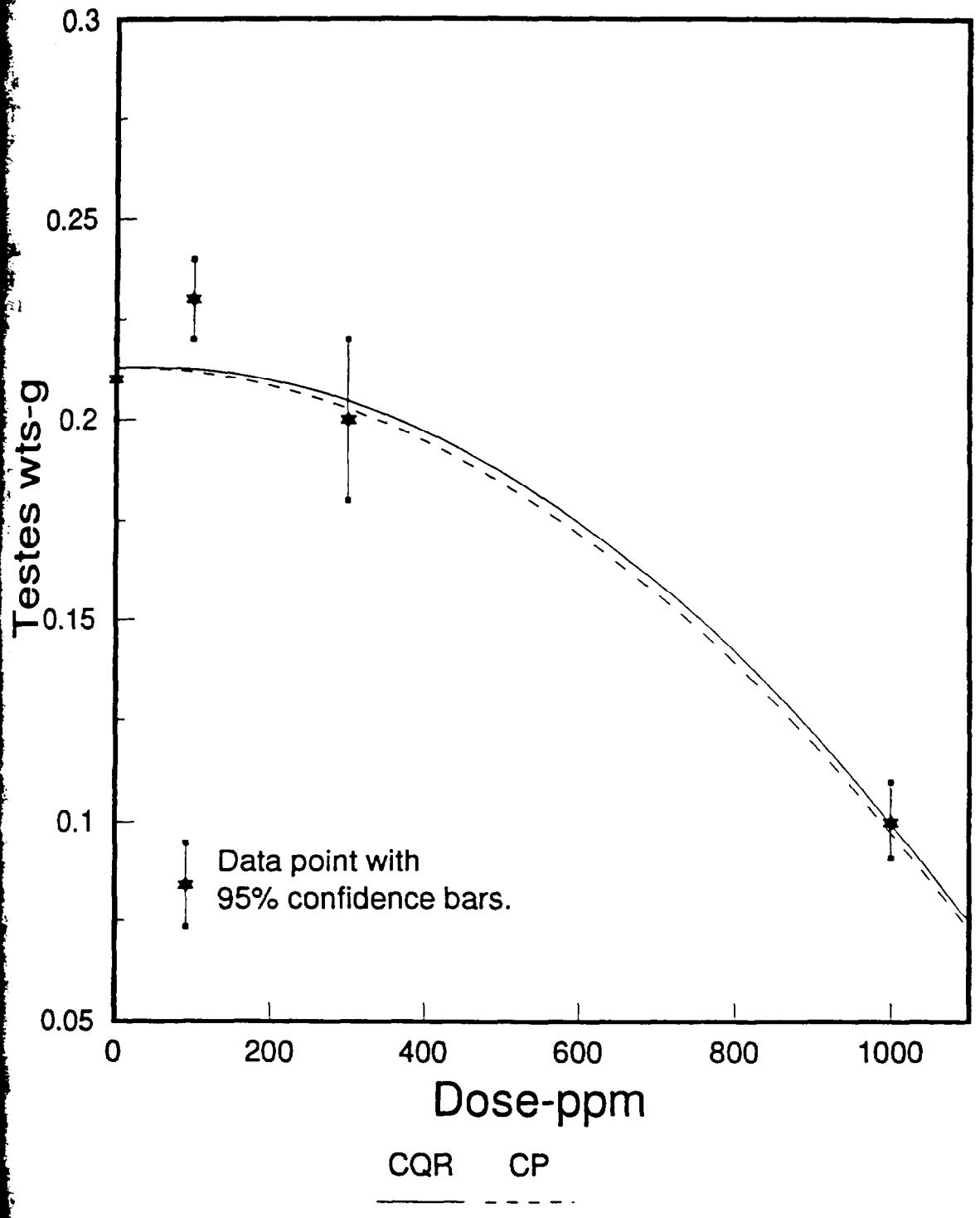
Source: Miller et al., 1981.

Figure 9. Testes Weights in Rats Following EGME Exposure



Source: Miller et al., 1981.

Figure 10. Testes Weights in Mice Following EGME Exposure



Source: Miller et al., 1981.

The estimated BMDs corresponding to a 5 percent relative decrease in testes weight in rats were 315 and 184 ppm, respectively, for the CQR and CP models. These two BMDs bracket the NOAEL of 300 ppm.

3.4.2. Additional Research

Additional research is needed to develop guidelines for handling issues related to lack of fit of BMD models. Ultimately, some decisions regarding suitable options when there is lack of fit will have to be developed. A particular need is guidance related to the biological and toxicological considerations that may influence decisions about dropping of doses.

As discussed earlier, use of estimates of internal dose at the site of toxicity could result in more appropriate RfDs, regardless of whether the NOAEL or BMD approach is used. It is recommended that more of the pharmacokinetic data needed for this purpose be generated, and that the experience gained from applying pharmacokinetic methods be used to develop guidelines for application of pharmacokinetic data in calculating RfDs. This effort is relevant to the issue of lack of fit because, as mentioned above, certain pharmacokinetic behaviors might account for dose-response patterns that are not strictly monotone (e.g., plateaus in response rates due to saturation of crucial metabolic pathways).

3.5. Measure of Altered Response

Another decision that must be made is the selection of a quantitative measure of altered response. Different types of measures are appropriate depending upon whether the response is in a quantal or continuous format.

For quantal data, two measures of increased response, "additional risk" and "extra risk," have been proposed (Crump, 1984). Additional risk is defined as

$$AR(d) = P(d) - P(0),$$

and extra risk as

$$ER(d) = [P(d) - P(0)] / [1 - P(0)].$$

In these equations, P(d) is the probability of response at dose d and P(0) is the probability of response in the absence of exposure (d = 0).

Additional risk is the additional proportion of total animals that respond in the presence of the dose. Extra risk is the fraction of animals that would respond when exposed to a dose, d, among animals who otherwise would not respond. Extra risk is typically used by EPA in risk assessments for cancer (Anderson et al., 1983).

Extra risk is additional risk divided by the proportion of animals that will not respond in the absence of exposure. Thus, extra risk and additional risk will coincide for responses that do not occur spontaneously.

Additional risk and extra risk differ quantitatively in the manner in which they incorporate background response. For example, if a dose increases one type of response from 0 percent to 1 percent and increases a second type of response from 90 percent to 91 percent, the additional risk is 1 percent in both cases. However, the extra risk is 1 percent in the former case and 10 percent in the latter case.

For continuous data, Crump (1984) suggested two measures of increased response analogous to those defined above for quantal data. The first is the difference between the mean response expected under exposure to dose d and the mean response expected in the absence of exposure:

$$|m(d) - m(0)|,$$

where m(d) is the mean value of the continuous measure of response for dose d. The vertical lines are symbols for absolute value and are incorporated to allow the expression to be applicable regardless of whether increases or decreases in the mean response are considered adverse.

The second measure proposed by Crump (1984) for continuous data normalized differences in mean responses by the background mean response:

$$|m(d) - m(0)| \; / \; m(0).$$

This measure of adverse response involves the fractional change in response rather than the absolute amount of change.

Crump (1984) also suggested that changes in a continuous endpoint could be assessed relative to the variability of that endpoint. His suggestion was to measure adverse response by

$$|m(d) - m(0)| \; / \; \sigma(0),$$

where $\sigma(0)$ is the standard error of the responses in the control group.

None of the measures proposed for continuous variables take into consideration the definition of an adverse effect (e.g., ranges of a continuous variable indicative of abnormality). Gaylor and Slikker (1990) suggested an approach for continuous data that would allow one to estimate the probability of an adverse effect from continuous data without the necessity of *first* categorizing the continuous responses observed (although it would still be necessary to

conceptualize a categorization into normal and abnormal ranges of response). Suppose there is a value of the response, A, that defines an adverse effect (e.g., responses greater than A are considered to be adverse). The approach of Gaylor and Slikker calls for dose-response modeling of the continuous data, followed by conversion of the mean and variance estimates to statements about the probability of observing adverse effects (e.g., effects greater than A) at given dose levels.

To implement the approach, one first fits a dose-response model to the observed continuous endpoints, and obtains estimates of the mean value of the response at a dose d, m(d), and the standard deviation for the observations at that dose, $\sigma(d)$. Then the probability, P(d), of an adverse response at dose, d, can be computed as

$$P(d) = \text{Probability } (\text{RESPONSE} \geq A).$$

This probability can be computed from knowledge of the mean m(d) and standard deviation, $\sigma(d)$. Using these probabilities, the equations for additional risk, AR, or extra risk, ER, for quantal responses can be applied in the subsequent steps of the BMD approach.

In order to use the approach suggested by Gaylor and Slikker (1990), one must assume a normal distribution for the continuous endpoints.[10] The need to assume some specific distribution is not a disadvantage compared to the other approaches to estimating risk for continuous responses, because a distribution must be assumed whenever a model is fit to continuous data (see Appendix A). An advantage of this approach is that it allows a common measure of adverse response to be used with both quantal and continuous data. Another advantage is that, unlike the data needed to define categorical responses from continuous data, the data necessary for implementation of this approach are likely to be summarized in a published report.

3.5.1. Examples

In examples presented in Sections 3.3.1 or 3.4.1 that used quantal responses (see Tables 3 and 5), extra risk was the measure of altered response used for BMD calculation. In the example of EGME-induced testicular toxicity (Table 6), for which responses were measured on a continuous scale, the measure of altered response used was relative change in weight (absolute change in mean testes weight normalized by the mean background—control—testes weight).

Consider the case of maternal effects induced by sulfamethazine during pregnancy. As part of a developmental toxicity study of sulfamethazine, the National Center for Toxicological Research (NCTR) conducted a preliminary study to determine the toxicity of that compound to pregnant animals (NCTR, 1981). Sulfamethazine was administered to CD rats at seven dose

---

[10]The method could readily be generalized to a non-normal distribution by replacing m(d) and $\sigma(d)$ by the parameters of that distribution. However, the data needed for efficient estimation of the parameters of a non-normal distribution generally will not be summarized in a published report of a study.

levels on gestation days 6 through 15. The maternal weight gain data for the entire gestational period are shown in Table 7. Weight gains were decreased at the three highest doses. Weight gains in the four lowest dosed groups, though larger than in controls, were not statistically different from controls. NCTR reported a significant trend for decreased weight gain, as tested by Jonckheere's test. Application of a procedure for determining trends for continuous endpoints based on the CP model (see Appendix A) established 600 mg/kg/day as the NOAEL. Both the CQR and CP models fit the data very well (Figure 11). The BMDs estimated from these models are displayed in Table 8. Shown in Table 8 are BMD estimates for two dose-response models and two measures of altered response (as well as three levels for the BMR and three confidence limit sizes; these are discussed below).

For both the CQR and CP models, the estimate of the BMDs depended greatly on the measure of risk. The differences across the two measures of risk were greater for the CP model (especially for smaller BMRs and for the larger confidence limits).

The results for the two models were most comparable when the absolute differences in the means were normalized by the background mean (and when either the BMR was 5 percent or greater or the confidence limit size was less than or equal to 95 percent). Normalizing by background response rates will enhance model independence.

3.5.2. Additional Research

Additional research is required to provide guidance regarding the measure of altered response that is most appropriate in particular instances. It is not clear when measures expressed relative to background (e.g., extra risk and absolute differences in means divided by background means) are preferable to measures expressed as absolute changes.

The method described by Gaylor and Slikker (1990) permits a BMD to be calculated from response probabilities irrespective of whether the underlying data are quantal or continuous. Although the method is conceptually sound, the statistical methodology needed for calculating confidence limits needs to be presented and computer software to implement the methodology needs to be developed. Support for these implementations and investigations of properties of the approach is needed. Particular aspects of the method that need to be addressed include questions regarding the definition of normal and abnormal ranges (whether based on professional, toxicological judgment or defined in terms of variability in the control—or other background—populations). Also of particular importance are methods for determining probabilities of being abnormal that are based on confidence limits rather than maximum likelihood estimates.

3.6. Selection of a Benchmark Level of Risk

The BMD is a lower statistical confidence limit on the dose corresponding to a specified level of risk called the benchmark risk, or BMR. Thus, before calculating a BMD, the BMR must first be specified. Several considerations may influence the selection of a BMR.

The first consideration is that the BMD is intended to replace the NOAEL, the largest experimental dose for which no statistically significant effects were observed. This suggests that

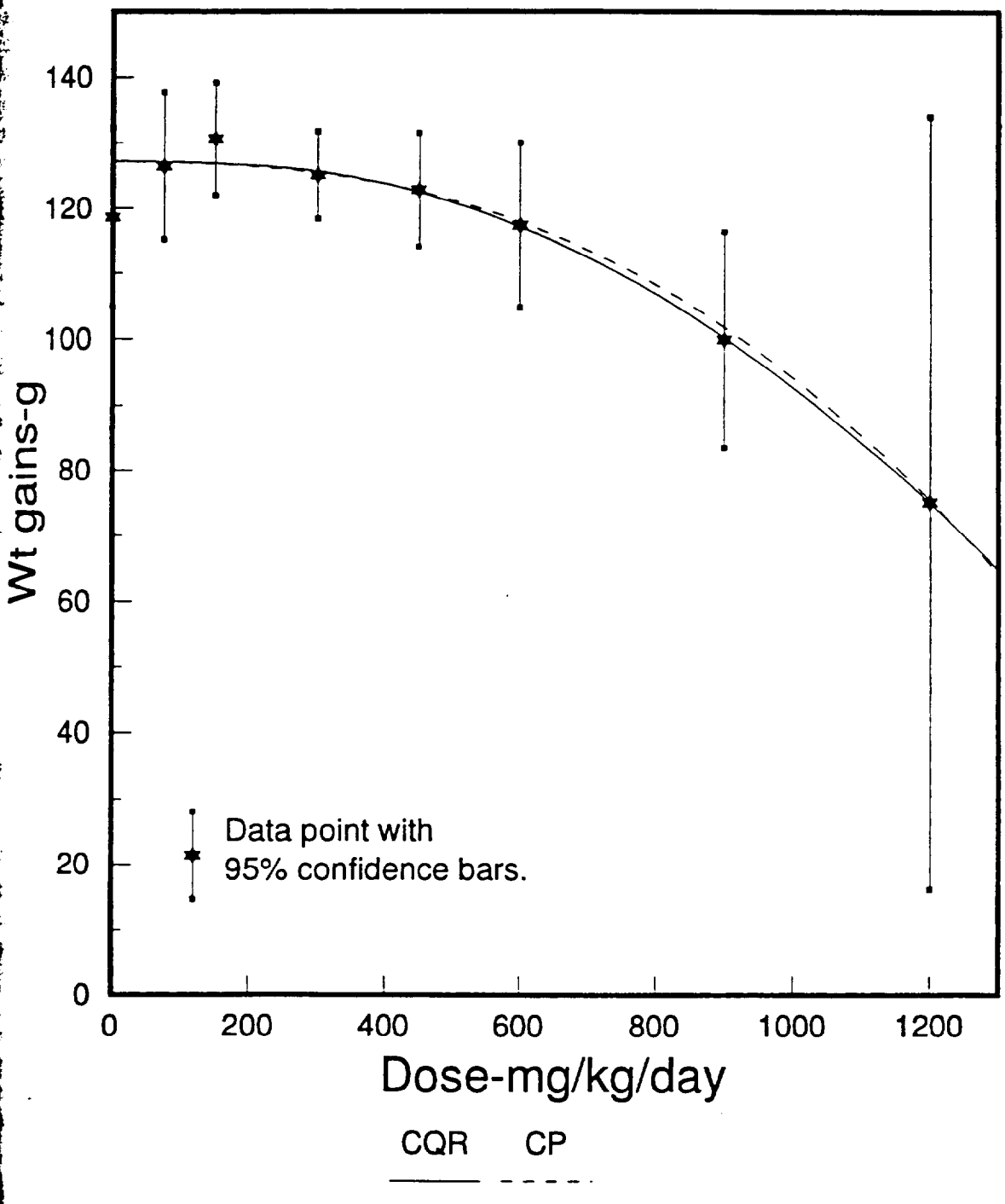
Table 7. Gestational Weight Gains in Pregnant Rats

Dose (mg/kg/day)	Average Weight Gain	SD	N
0	118.6	24.7	13
75	126.4	14.8	7
150	130.6	10.5	6
300	125.1	8.2	6
450	122.8	10.6	6
600 (NOAEL)	117.4	14.1	5
900	100.0	20.1	6
1200	75.2	58.9	4

SD = standard deviation
N = number of pregnant animals for which weight gains were determined

Source: NCTR, 1981.

Figure 11. **Weight Gain during Gestation in Rats Exposed to Sulfamethazine**



Source: NCTR, 1981.

Table 8. BMDs (mg/kg/day) Calculated for Sulfamethazine Data

Model	Measure of Risk	BMR	Confidence Limit Size 90%	95%	99%
CQR ($p = 0.73$)[a]	Absolute difference of means	10%	49.0	47.3	44.6
		5%	34.7	33.5	31.5
		1%	15.5	15.0	14.1
	Absolute difference of means normalized by background mean	10%	558	540	510
		5%	395	382	361
		1%	176	171	161
CP ($p = 0.70$)[a]	Absolute difference of means	10%	28.9	18.0	5.29
		5%	19.0	11.2	2.82
		1%	7.18	3.71	0.655
	Absolute difference of means normalized by background mean	10%	533	491	405
		5%	355	311	226
		1%	136	105	54.5

[a] p-values for goodness-of-fit

the BMR should be selected near the low end of the range of increased risks that can be detected in a bioassay of typical size.

Another consideration is that an important goal of the BMD approach is that the approach be relatively model-independent; that is, different dose-response models that fit the data should give comparable estimates of the BMD. However, it is well known that different mathematical dose-response models can fit data equally well and yet produce widely divergentestimates of risk at doses far below the range that produce measurable increases in response (Crump, 1985). Thus, for the BMD approach to be relatively model-independent, the BMR cannot be much smaller than increased responses that can be measured reliably in experimental groups of typical size.

Some simple quantitative considerations can provide guidance with respect to the setting of the BMR. Consider a quantal response in a relatively large dose group of 100 animals and suppose that the observed response rate is 1 percent. A 95 percent confidence interval for the true rate of response ranges from 0.25 percent to 5.4 percent. (A confidence interval for the difference between the rate in this group and that in a control group would be even larger.) This illustrates the fact that increased responses of 1 percent or less cannot be measured with much precision in bioassays of typical size. That is, a BMR below 1 percent would be expected to be outside the range of risks that could be measured accurately in typical experiments.

Various papers (Crump, 1984; Dourson et al., 1985; Kimmel and Gaylor, 1988; Gaylor, 1989) have proposed a BMR for quantal responses in the range of 1 percent to 10 percent. Less attention has been given to corresponding levels for continuous effects. If the approach of Gaylor and Slikker (1990) (see Section 3.5) is used for continuous effects, then it may be possible to use the same BMR for continuous responses as for quantal responses.

3.6.1. Examples

In the example of EGPE-induced toxicity in the spleen of rats (Section 3.4.1, Table 5), BMDs were calculated for BMRs of 10 percent, 5 percent, and 1 percent. For the two quantal models examined, the BMD estimates differed by slightly more than a factor of 2 for 10 percent extra risk. There was less agreement at lower risk levels and at an extra risk of 1 percent, BMDs from the two models differed by a factor of 7.3. For the QQR model, the BMDs corresponding to 10 percent and 5 percent extra risk bracket the NOAEL. All of the BMDs calculated for the QW model fall below the NOAEL, with the BMD for 10 percent risk being about one-half the NOAEL value.

In the example of sulfamethazine-induced effects on the continuous variable, gestational weight gain (Table 7), BMDs were calculated for three levels of the BMR, 10 percent, 5 percent, and 1 percent (Table 8). For the two models considered, and for each of the measures of risk, the results were more similar across models (i.e., there was greater model independence) when the BMR was 5 percent or greater.

3.6.2. Additional Research

One of the desired features of the BMD approach is that, since extrapolation far beyond the range of the data is avoided, the procedure should be relatively independent of the dose-

response model utilized. The extent to which this is the case depends in part on the BMR selected. As lower BMRs are used, the corresponding BMDs should become more model-dependent because one is extrapolating further beyond the range of the data. This was observed in the examples. However, as observed in the examples, there will be some divergence in BMDs regardless of the BMR selected. The goal in selecting the BMR is to make it as small as practical without the BMD becoming too model-dependent. Although a BMR of 1 percent to 10 percent has been recommended by various authors (Crump, 1984; Kimmel and Gaylor, 1988; Gaylor, 1989), there has been no systematic study of data from a number of chemicals to determine how model-dependent the BMD is for various values of the BMR. Such a study could provide a more definitive basis for selection of a BMR and could evaluate the model uncertainty at the recommended BMR. It could also provide experience on the performance of various models and information on how well models fit data and what problems might arise from their application.

3.7. Confidence Limit Calculation

Decisions to be made in the calculation of a lower confidence limit for the dose corresponding to the BMR involve selection of the procedure for calculating confidence limits and the size of the confidence limits. Recall that the BMD is defined to be the lower confidence limit on dose corresponding to the BMR. The lower limit, as opposed to the maximum likelihood estimate, is used for several reasons, the foremost among them being the fact that statistical confidence limits account for the sample size of an experiment. The fact that NOAEL determinations do not account for sample sizes was one of the major criticisms of the NOAEL approach. Other factors that make the lower confidence limit preferable to the maximum likelihood estimate include the fact that the lower limit will be more stable to minor changes in the data and that the lower limit may be estimable even in some cases where the maximum likelihood estimate is not.

Confidence limits based on maximum likelihood theory have a number of desirable statistical properties (Cox and Lindley, 1974) and are typically preferred. Confidence limits based on this approach can utilize either the asymptotic distribution of the parameter estimates themselves or the asymptotic distribution of the likelihood ratio statistic (Cox and Lindley, 1974). Crump and Howe (1985) found that the latter approach (described in Appendix A) appeared to have superior statistical qualities in dose-response applications. This approach is incorporated into GLOBAL 82 (Howe and Crump, 1982), the computer program that has been used by EPA for dose-response modeling for cancer.

The size of statistical confidence limits ranges from 90 percent to 99 percent in most applications. Rather than being based on scientific rationale, this range seems to be purely conventional. EPA has generally employed one-sided 95 percent confidence limits in risk assessments for cancer effects (Anderson et al., 1983).

3.7.1 Example

In the example of sulfamethazine effect on weight gain during pregnancy (Section 3.5.1, Tables 7 and 8), BMDs were calculated for three sizes of confidence limits. For the CQR model, the choice of confidence limit size had very little impact on the BMD estimates. For the CP model, however, the choice of confidence limit size was much more important, especially

when absolute difference in the means was used as the measure of risk. The importance of confidence limit size with the CP model increased as BMR decreased; e.g., the BMD estimate for the 1 percent BMR was more sensitive to the choice of confidence limit size than was the estimate for the 5 percent BMR.

The results for the two models were most comparable when the absolute differences in the means were normalized by the background mean (and when either the BMR was 5 percent or greater or the confidence limit size was less than or equal to 95 percent). This suggests that not only will the BMD estimates be model-dependent for low levels of risk, but that they may also be model-dependent when wide confidence limits are calculated.

3.7.2. Additional Research

The appearance of the interactions discussed in the example highlights two features: the care with which one must consider the options for all of the decision points, and the need for additional research to investigate the interrelationships among the decisions. As an extension to the research suggested in Section 3.6.2, one should also consider the impact of the size of the confidence limits on the model independence of the BMD approach. It is clear that this cannot be done in isolation from the choices concerning the BMR level. Some guidelines for the selection of confidence limit size also could be developed that consider the adequacy (from a health-protective policy perspective) of confidence limits of various sizes.

3.8. Determination of a Single BMD

Depending on the options selected for choosing models and the responses to model, the procedures discussed to this point may yield a single BMD, multiple BMDs calculated from applying multiple models to individual responses, multiple BMDs calculated from different responses in a single study, and/or multiple BMDs calculated from different studies.

Multiple BMDs may arise when different models fit the data for a single response in a single study. Different BMDs could also come from a single study if more than one response is modeled. Selection of any BMD other than the smallest one from that study might lead to an RfD that is not protective against the effect corresponding to the smallest BMD.

Different BMDs could arise for the same response from different studies. Potential differences among studies with regard to species of animal studied, dosing patterns, and other features of experimental design make it difficult to specify a general rule that would be applicable in all situations.

3.8.1. Examples

In the examples discussed above, BMDs for a single endpoint in a single study have been calculated using two different models (Tables 3, 5, 6, and 8). Since it may not be possible to eliminate one model from consideration (either because of lack of fit, inappropriate statistical assumptions, or biological considerations; see Section 3.3) some judgment must be made regarding the treatment of the pairs of BMDs arising from the two models.

Consider the example presented in Table 3. Two options for dealing with multiple BMDs from a single endpoint can be illustrated. The first is to use the smallest of the BMDs, which in this case is 0.31 mg/kg/day. The second option is to combine the estimates. If a geometric average is used, the resulting BMD estimate for acrylamide-induced nerve degeneration is 0.51 mg/kg/day. For the sake of this example, attention is limited to the two models, QQR and QW.

3.8.2. Additional Research

Determining how to deal with multiple BMDs is an issue that requires more extensive Agency discussion. Perhaps examination of current RfD/RfC Workgroup policies could suggest guidelines for this issue.

3.9. Uncertainty Factors

Once a unique BMD is calculated, an RfD is obtained by dividing the BMD by one or more uncertainty factors. This same step is required in the NOAEL approach, but the uncertainty factors are applied to the NOAEL rather than the BMD.

The uncertainty factors that used to be routinely applied to NOAELs (Table 1) have been criticized as being arbitrary. However, they do have a history of use; this has allowed a sense of their utility to develop, encouraging persons in the field to become familiar and perhaps comfortable with their use. No comparable experience exists for application of uncertainty factors to BMDs. Before adopting uncertainty factors for BMDs, it might be useful to compare BMDs with NOAELs for a variety of substances. This would permit the relative magnitudes of BMDs and NOAELs to be compared and could suggest uncertainty factors that would be appropriate for BMDs.

New approaches to the definition and calculation of uncertainty factors are being investigated and may provide new rationales for uncertainty factors (Hattis and Lewis, 1992). This work should be applicable to BMDs as well as to NOAELs. However, it should be noted that, unlike the NOAEL, the calculation of the BMD depends on the BMR as well as the size of the statistical confidence bound employed. These additional considerations may need to be accounted for when selecting uncertainty factors for BMDs.

Other factors that conceivably could affect the selection of uncertainty factors include the slope of the dose-response curve and some biological considerations (e.g., relating to the possibility of a threshold for the responses under investigation). The manner in which these factors should affect uncertainty factors is unclear at present.

Another option for selection of uncertainty factors has been presented by Kimmel and Gaylor (1988). In this option, the selection of uncertainty factors is tied to the specific level of extra risk (e.g., 10^{-5}) that is deemed to be sufficiently health-protective. If, for example, the BMD is calculated for a risk level of 1 percent (10^{-2}), and if a risk of 10^{-5} is considered to be adequate for an RfD risk level, then the uncertainty factor would be $10^{-2}/10^{-5} = 1{,}000$. That is, the RfD is obtained from the BMD by dividing by an uncertainty factor of 1,000. In general, this option calls for an uncertainty factor equal to the ratio of the BMR and the risk level corresponding to the presumed human safe dose.

This option is equivalent to extrapolating to the risk level corresponding to the presumed human safe dose with a linear dose-response function (e.g., the linearized multistage approach that EPA applied to cancer data [Anderson et al., 1983]). If the true dose-response model was linear, the resulting RfD would be a true 95 percent lower limit on the dose corresponding to an extra risk of 10^{-5}. However, if the dose response is highly nonlinear (e.g., thresholdlike), as is considered possible for many non-cancer effects, this option could result in a highly conservative RfD. Moreover, this approach introduces a different philosophy than that underlying the NOAEL and other BMD approaches, which is to account for uncertainty in low-dose response using uncertainty factors based on toxicological judgment, rather than by specifying a conservative dose-response model.

3.9.1. Example

Consider the example in Section 3.3.1 of acrylamide neurotoxicity (Table 3). If the same uncertainty factors as used in the NOAEL approach are considered appropriate, the factors that might be relevant are a factor of 10 for animal-to-human extrapolation and another factor of 10 for human variability, for a total uncertainty factor of 100. Application of this uncertainty factor to the two BMD estimates shown in Table 3 yields RfDs of 3.1 μg/kg/day or 8.3 μg/kg/day. If an average of the two BMDs were selected (see the example in Section 3.8.1), then the resulting RfD would be 5.1 μg/kg/day.

If the option for uncertainty factor selection described in Kimmel and Gaylor (1988) were to be used, one must determine a level of risk that is sufficiently low to be considered acceptable for human populations. Let us suppose that for the endpoints under consideration in this example, a risk level of 10^{-4} is acceptable. Then, the uncertainty factor that is applied to the BMDs is determined by the ratio of the BMR (in this case 0.05) and the acceptable level of risk. Thus, the uncertainty factor selected in this case would be $0.05/10^{-4} = 500$. The RfDs calculated using this approach would be 0.62 μg/kg/day and 1.7 μg/kg/day (for the BMDs in Table 3) or 1.0 μg/kg/day (if the average of those BMDs were used).

The NOAEL derived from these data was 0.5 mg/kg/day.[11] The typical uncertainty factor applied to that NOAEL, a factor of 100, yields an RfD of 5.0 μg/kg/day. That value is very close to the RfD calculated using the average BMD and the same uncertainty factor as the NOAEL approach. The RfDs calculated from the BMDs using the Kimmel and Gaylor (1988) approach to deriving uncertainty factors were about three to eight times smaller than the RfD based on the NOAEL.

---

[11]Although Johnson et al. (1986) reported that the high-dose group experienced significantly greater mortality than the controls, the data reported in the manuscript are not adequate for conducting a mortality adjusted test. However, the authors noted that the Mantel-Haenszel test showed a significant dose-related trend in degeneration of tibial nerves when applied to all the dose groups, and they stated that the degeneration results for doses of 0.5 mg/kg/day and below were "comparable to controls." The Mantel-Haenszel test applied to the data without adjustment for survival differences was not significant when the highest dose group was ignored. From such information, we conclude that 0.5 was the NOAEL for tibial nerve degeneration in male rats.

3.9.2. Additional Research

The uncertainty factors applied to a NOAEL to calculate an RfD have been applied extensively for a number of years. Although the "traditional" factors (Table 1) are not firmly based on objective criteria, they were developed after deliberation and debate by toxicologists, and consequently reflect informed judgment as to the degree of safety afforded by different uncertainty factors. A public process such as this has not yet been used to determine uncertainty factors for use with BMDs.

It is recommended that a panel of toxicologists and scientists from related fields develop recommendations concerning uncertainty factors to use with BMDs. The panel could consider whether the use of the BMD approach increases or decreases the recommended RfD in general, as well as possibly identify exceptional cases for which the BMD approach may not be recommended. Since the RfD depends upon the BMR used and the size of the confidence interval, as well as the uncertainty factors applied, the panel should also review the recommendations concerning those factors.

The roles of the dose-response slope and of biological considerations (e.g., the likelihood of thresholds) also could be explored by the panel. Such a panel would require some background data. It is recommended that, as a basis for such deliberations, the BMD procedure be applied to data from a variety of chemicals and toxic endpoints and that results be compared with those obtained by applying the NOAEL approach.

Such an investigation of uncertainty factors in the context of the BMD approach would complement the recent work that has been undertaken to reconsider the basis of the uncertainty factors used in the NOAEL approach.

3.10. Summary of BMD Decisions

The decisions required in the implementation of the BMD approach have been presented above. Some of the available options for each of the decisions, including options that have been proposed in the literature, also have been presented. Options for each of the decisions are summarized in Table 9. By no means do these exhaust all the possibilities. The options presented were selected because they were judged to have scientific merit, seemed reasonable, and/or have a history of use.

4. DETAILED COMPARISON OF NOAEL AND BMD APPROACHES

4.1. Conceptual Basis

A NOAEL for an experiment (if one exists) is an experimentally determined exposure level at which there is no statistically or biologically significant increase in the frequency or severity of adverse effects between the exposed population and its appropriate control; some effects may be produced at this level, but they are not considered adverse nor precursors to adverse effects. In an experiment with several NOAELs, the regulatory focus is primarily on the highest one, leading to the common usage of the term NOAEL as the highest exposure without adverse effect (U.S. EPA, 1991). The NOAEL has sometimes been referred to as an "experimental threshold," although it should not necessarily be considered an estimator of a

Table 9. Summary of Decisions and Options for BMD Approach

Decision	Options
1. Experiments to include	a. All relevant, high-quality studies b. A single, "critical" study
2. Responses to model	a. All responses from selected studies b. Responses observed at LOAEL
3. Format of data	a. Convert continuous data to categorical data b. Transform continuous data (e.g., log-transformation) c. Retain original, continuous format
4. Mathematical model(s)	a. All models with adequate fit to the data b. Models with most appropriate statistical assumptions c. Models most appropriately reflecting biological considerations (e.g., threshold) d. Models satisfying combinations of a-c
5. Handling lack of fit	a. Try more flexible model(s) b. Omit high-dose data if lack of fit is due to those data c. Use measure of internal dose
6. Measure of altered response	
Quantal data	a. Additional risk b. Extra risk
Continuous data	a. Absolute difference in means b. Absolute difference in means normalized by background mean c. Absolute difference in means normalized by background standard error d. Gaylor and Slikker approach with additional risk e. Gaylor and Slikker approach with extra risk
7. BMR definition	a. 1% to 10% risk

Table 9. Summary of Decisions and Options for BMD Approach (cont.)

Decision	Options
8. Confidence limit calculation	
Method	a. Likelihood theory, based on asymptotic distribution of likelihood ratio statistic b. Likelihood theory, based on asymptotic distribution of parameter estimates
Size	90% to 99%
9. Specific BMD for RfD calculation	
Multiple BMDs for a single endpoint	a. Select smallest BMD b. Combine BMDs (e.g., geometric average)
Multiple BMDs from a single study	a. Select smallest BMD
Multiple BMDs from multiple studies	a. Select smallest BMD b. Average BMDs for different species and/or sexes c. Use most appropriate species and/or sex
10. Uncertainty factors	a. Use same factors as used in NOAEL approach b. Use NOAEL factors modified by average ratio, BMD/NOAEL c. Use risk-based factors (Kimmel and Gaylor, 1988) d. Use factors dependent on choice of BMR and confidence limit size e. Use factors that consider dose-response slope and/or biological considerations

biological threshold (if one exists). The definition of the NOAEL implies that it is the highest experimental dose just smaller than the LOAEL. It is also the case that a NOAEL represents a dose at which there is no significant change (from control) in response. There may, in fact, be some instances where adverse effects are seen at the NOAEL, but not at a level that is statistically significant.

The NOAEL approach traditionally has been used for effects that are expected to have a threshold. On the other hand, use of mathematical dose-response models has generally been reserved for effects, particularly cancer effects, that are considered not to have a threshold. Conceptually, there is no reason why mathematical dose-response models cannot be applied to threshold effects as well as non-threshold effects. A threshold can be incorporated into a model as a parameter, and the value of the threshold can be estimated. In fact, several of the dose-response models listed in Table 4 for use in the BMD approach explicitly incorporate a threshold dose, d_0 (QLR, QQR, CLR, and CQR).

Further, when calculating a BMD using a dose-response model, it is not strictly necessary that threshold effects be modeled with threshold models and non-threshold effects with non-threshold models. (This is fortuitous since the existence or non-existence of a threshold is generally not known with certainty.) This is because in the calculation of a BMD, the mathematical model is used only to estimate doses corresponding to a given level of increased response (the BMR). Thus, even if a threshold does exist for an effect, the dose-response model is used for prediction only at doses above the threshold.

4.2. Relative Sizes of NOAELs and BMDs

The fact that a BMD corresponds to a specified level of adverse change in response (for quantal data, generally 1 percent to 10 percent increased risk, as discussed earlier) and a NOAEL ostensibly corresponds to no increased risk does not imply that NOAELs will necessarily be smaller than BMDs (and consequently that larger uncertainty factors may be appropriate for BMDs). First, a BMD is defined as a statistical lower limit, which introduces an element of conservatism in its definition. Second, one cannot conclude that no adverse effects are possible at a NOAEL. The BMD corresponding to an extra risk of 1 percent was smaller than the corresponding NOAEL for each of ten data sets studied by Gaylor (1989). Among five sets of quantal data studied by Crump (1984), the BMD corresponding to an extra risk of 1 percent was larger than the NOAEL in one case by a factor of 1.4, and smaller than the NOAEL in three cases by factors ranging from 1.1 to 2.6 (one data set did not define a NOAEL). However, it is unclear whether the data sets used in these studies are typical of those to which the BMD method would need to be applied if the method is used routinely.

4.3. Constraints Imposed by the Experimental Design

Whereas the BMD can theoretically assume any value, the NOAEL is constrained to be one of the experimental doses. This constraint can appear unnecessarily restrictive in some cases. If, for example, only a marginally significant effect is seen at the LOAEL, and there is a large gap between the LOAEL and the next lowest dose, then the estimated NOAEL could be considerably smaller than would be obtained from a study employing more doses or a more judicious selection of doses. On the other hand, a BMD could be estimated at a dose intermediate between the LOAEL and the NOAEL.

The NOAEL approach must be modified whenever effects are seen at all doses and consequently a NOAEL is not determined. Two approaches have been used in this situation. One approach has been to require the study to be repeated at lower doses in order to define a NOAEL. This alternative may be costly and time-consuming, and may appear to be unnecessary whenever a clear dose response is defined by the original experiment (Crump, 1984). The other approach has been to use the LOAEL instead of a NOAEL in calculating the RfD, but require an additional uncertainty factor of 10 to be applied (see Table 1). This approach appears to be ad hoc, particularly since the size of the recommended uncertainty factor does not depend upon the level of effect seen at the LOAEL. On the other hand, the BMD approach does not have this limitation since a BMD can be determined regardless of whether a NOAEL is defined by the data.

4.4. Number of Experimental Subjects and Their Distribution into Treatment Groups

One of the major differences between the NOAEL and BMR approaches is the manner in which they incorporate sample size. If fewer animals are tested per group, it is less likely that a real difference in response rates between two groups will be detected. Thus, experiments with fewer animals per dose group will tend to find larger NOAELs than experiments with more animals per dose group. These considerations have led EPA to impose minimum requirements for numbers of animals per test group. For example, the guidelines for developmental toxicity testing protocols recommend at least 20 animals per dose group (U.S. EPA, 1986). This aspect of the NOAEL approach is the opposite of what would seem appropriate; a larger study should afford greater evidence of safety and therefore should result in a larger RfD.

On the other hand, a BMD will appropriately tend to be larger when estimated from a study employing larger numbers of animals per dose group. This is because a BMD is defined as a lower statistical confidence limit and a larger study will tend to define narrower confidence bounds (i.e., larger lower limits and smaller upper limits).

With either the NOAEL or the BMD approach it is desirable to have data from several treatment groups. With the BMD approach such data help to define the shape of the dose response, which is estimated by the model; consequently, such data permit more accurate estimation of the BMD. Having several treatment groups is also desirable when applying the NOAEL approach since this increases the range of possibilities for the NOAEL and consequently may increase the precision of the NOAEL approach.

For a given total number of experimental animals, the more dose groups in the experiment, the fewer animals that can be tested at each dose. Dividing a given total number of animals into more treatment groups will generally not have a major impact upon a BMD calculation because the BMD approach does not focus on dose groups individually, but instead fits a single dose-response model to all of the available data from a study. This is not the case with the NOAEL approach, however. Since this approach compares individual responses at individual doses to responses in a control group, dividing a given number of animals into more

groups will decrease the power for detecting an effect at any particular dose, and consequently tend to result in a higher NOAEL.[12]

4.5. Incorporation of Dose-Response Information

A NOAEL may be based solely on information concerning whether an effect is observed at particular doses; the relationships among the magnitudes of the responses at the given doses may not be taken into account. On the other hand, the BMD is based on a dose-response curve that naturally takes into account the shape of the dose response.

This is illustrated in Figure 12 in which the QW model has been used to determine BMDs for two hypothetical data sets. The first data set (marked by x's) has a steep dose response above the LOAEL, which in this example equals 1 mg/kg/day. The second data set (marked by o's) has identical responses up to the LOAEL, but then has a more gradual dose response at doses above the LOAEL. Also plotted are BMDs for the two data sets corresponding to risks of 1 percent. The first data set produces a higher BMD than the second, which seems reasonable given the respective dose-response shapes. On the other hand, the NOAEL, which is insensitive to the steepness of the dose response, is the same for both data sets.

4.6. Sensitivity to Data Interpretation and to Small Changes in Data

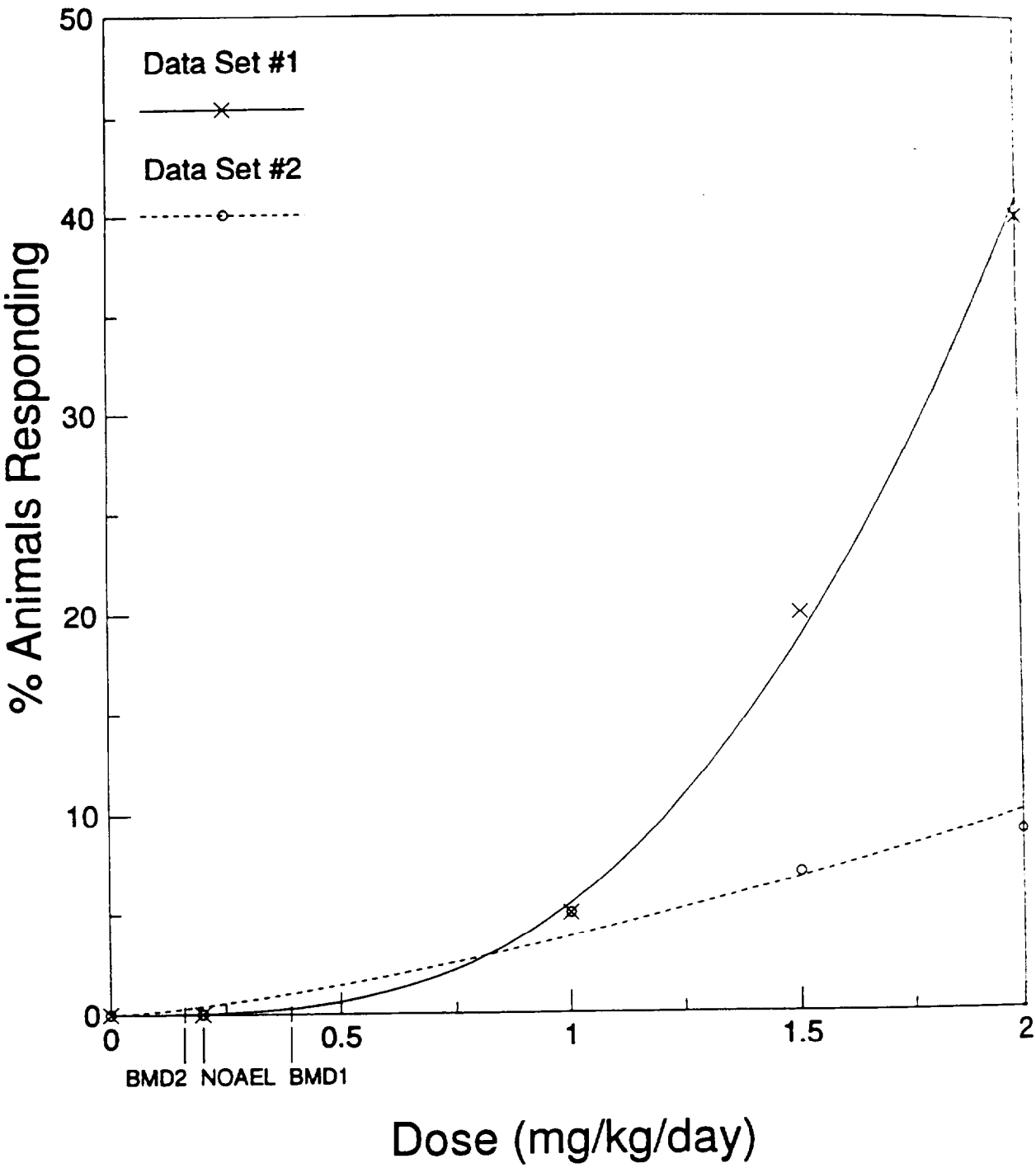
The NOAEL approach involves a number of decision points for which slight changes in data can have a sizable effect on the outcome. Determination of a LOAEL and a NOAEL are based, at least in part, upon the degree of statistical significance, and in marginal situations changes in responses of only a few animals (or in even a single animal) can change a significant response to nonsignificant and vice versa. Further, according to the definition of a NOAEL, effects that are not statistically significant can be determined to be biologically significant. For example, if an effect is found in several animals but it is rare to observe the effect spontaneously in untreated animals, it might be considered to be related to dose even if the rate of response is not formally statistically significantly different from that in controls. In both of these situations, small changes in data or even differences in interpretation of data could have a substantial effect upon the NOAEL and consequently upon the RfD. On the other hand, the calculation of BMD does not require judgments about whether an effect is present in individual dose groups. The BMD also appears to be less sensitive than the NOAEL to small changes in the data.[13]

---

[12]This effect could be mitigated by using a statistical trend test to test for a dose response trend among the doses at and below a potential NOAEL. Such a test utilizes data from all of the doses in a range rather than comparing a single dose group to a control group. The NOSTASOT (no statistical significance of trend; Tukey et al., 1985) test procedure was proposed specifically for this situation.

[13]A situation in which a BMD might be affected by a small change in data is when there is a borderline lack of fit of models to the data and a decision must be made regarding whether to omit data at the highest dose.

Figure 12. **Example of BMDs Calculated from Steep versus Gradual Dose Responses**



4.7. Model Sensitivity

Since the NOAEL approach does not involve the use of dose-response models, the issue of model sensitivity applies only to the BMD approach. Since the calculation of a BMD does not involve extrapolation of results to doses far below those for which effects are observed, the BMD approach has been presented as being relatively model-independent (Crump, 1984). However it appears that this issue has not been investigated thoroughly. Crump (1984) applied four dose-response models to each of four sets of quantal data and one set of continuous data. The ratios of the largest to the smallest of the four BMD_5s for each of the five data sets were 1.2, 1.1, 1.2, 1.4, and 1.3. The corresponding ratios for the BMD_{10}s were 1.3, 1.1, 1.2, 1.2, and 1.1. These ratios are small compared to the large model differences that occur when extrapolating to much lower doses (Crump, 1985).

4.8. Quantitative Estimates of Risk

The NOAEL represents a dose at which there is no significant increase in response. However, it cannot be assumed to represent a no-effect level in a large population (i.e, a population threshold). The risk at the NOAEL has been estimated as being generally as high as about 5 percent (Gaylor, 1989).

Unlike the NOAEL approach, the BMD approach associates a risk with each dose based on a mathematical dose-response model. The calculation of a BMD utilizes the predictions of the model only at doses at and above the BMR (doses that typically correspond to altered responses of 1 percent or greater). However, if desired, the model used to calculate the BMD also could be used to estimate risks for lower doses, even though this is not part of the BMD approach, per se. If such low-dose extrapolation is performed, it should be recognized that the results are likely to be highly model-dependent.

There are two simple ways this extrapolation could be carried out. The first would be to simply utilize the predictions of the model used in the calculation of the BMD, or confidence limits on these predictions. The second method, which is similar to the method proposed by Kimmel and Gaylor (1988) for determining uncertainty factors for the BMD approach (see Section 3.9 on uncertainty factors), is to assume that the dose response is linear below the BMR. Thus, the risk at a dose, d, which is less than the BMR would be estimated as (d/BMD)*BMR. This approach would generally yield results similar to those obtained by applying the linearized multistage approach that EPA applies to carcinogens (Aynderson et al., 1983). This approach will greatly overestimate the risk from substances with a dose response that includes a threshold or is highly non-linear. Even so, such conservative estimates of risk could be useful in some applications. For example, if such a conservative procedure predicted a low risk, this would indicate that the true risk is at least this low and possibly much lower.

4.9. Statistical Expertise

Both the NOAEL and BMD approaches require the use of statistical methods. With the NOAEL approach, statistical tests for comparing two groups of data as well as tests for a dose-response trend across several dose groups may be needed. These same tests may be required in applying the BMD approach (e.g., to determine the critical study). In addition, the BMD method requires statistical methods to fit mathematical dose-response models to data.

Statistical goodness-of-fit tests are needed to determine how well these models describe the data. Further, a statistical confidence limit on dose corresponding to a given BMR needs to be calculated in order to define the BMD. Thus, the BMD method definitely requires greater use of statistical methodology than the NOAEL approach.

Existing computer packages can perform all of the statistical tests required. Moreover, programs are available that fit most of the models listed in Table 4 to data using the method of maximum likelihood (Crump, 1984). Those programs also test goodness-of-fit and calculate the required confidence intervals. Consequently, a person with scientific credentials who understands basic statistical concepts and the basic ideas of the NOAEL and BMD approaches and who has access to the necessary computer programs and facilities for running them should be able to perform the necessary analyses. Although a statistician should not be required to perform the calculations, one should be available for consultation. Implementation of these methods would be facilitated if a user-friendly, special-purpose program was available that could perform all of the necessary calculations. Also useful would be some special training (e.g., a 1-day seminar) for presentation of the statistical methods used in the BMD approach and the use of computer programs for making the necessary calculations.

5. SUMMARY OF RESEARCH NEEDS

The discussions above suggested several areas in which additional research into the BMD approach could be of value. These are summarized here. Two additional investigations/developments are also discussed. Several of these research needs could be addressed through a study that involves computing BMDs corresponding to various BMRs using several dose-response models for a number of data sets.

5.1. Summary of Research Needs Related to BMD Decision Points

The areas identified in the preceding material that require additional research are the following:

1. Development of dose-response models and related methods for use with various types of data (see Section 3.3.2).

2. Guidelines for handling lack of fit (Section 3.4.2).

3. Development of methods of applying pharmacokinetic considerations (Section 3.3.2).

4. Guidelines for selecting appropriate measure(s) of altered response (Section 3.5.2).

5. Study of the sensitivity of the BMD to choice of model, particularly in relation to the level of the BMR (Section 3.6.2) and to the confidence limit size (Section 3.7.2).

6. Guidelines for selecting a single BMD when more than one is calculated (Section 3.8.2).

7. Investigation of uncertainty factors (Section 3.9.2).

5.2. Additional Topics for Investigation/Development

5.2.1. Comparison of Dose-Response Curves for Different Types of Data and Toxic Endpoints

In the process of applying the BMD approach to a number of data sets, as is required for the last two research recommendations above, it could be worthwhile from a theoretical perspective to evaluate the various dose-response curve shapes for different forms of data (e.g., quantal versus continuous), for different toxic endpoints, and for different chemical classes. Such a study could provide information on which endpoints appear to be have a threshold response versus a non-threshold response and whether the dose responses of the same effect from different chemicals appear to have the same shape. This information could be used to construct hypotheses regarding underlying mechanisms that could be tested in subsequent experiments. It would be particularly interesting to determine whether non-cancer responses appear in general to be more "thresholdlike" than cancer. This research would have implications concerning the appropriateness of applying different types of procedures for setting allowable exposure for carcinogenic effects and various categories of non-carcinogenic effects.

One way to conduct such a study would be to apply the QW and CP models and study the values of the shape parameter, k, from these models. A value of $k = 1$ is consistent with a linear no-threshold dose-response, whereas large values of k are more indicative of a threshold.

5.2.2. Development of User-Friendly Computer Programs for Calculating RfDs

Having a single computer program available for performing the statistical calculations needed for testing hypotheses and fitting models could facilitate routine implementation of the BMD approach and also aid in implementing the NOAEL approach. A user-friendly program designed explicitly for this purpose could be particularly helpful for non-statisticians. Such a program might also incorporate elements of an expert system that guides the user step-by-step through all of the stages of both the BMD and NOAEL approaches, beginning with the data review and the decisions necessary to select the critical study.

6. REFERENCES

Andersen, M.; Clewell, H.; Gargas, M.; Smith, F.; Reitz, R. (1987) Physiologically based pharmacokinetics and the risk assessment process for methylene chloride. *Toxicol. Appl. Pharmacol.* 87:185-205.

Anderson, E.; Carcinogen Assessment Group of the U.S. Environmental Protection Agency. (1983) Quantitative approaches in use to assess cancer risk. *Risk Anal.* 3:277-295.

Bickel P.; Doksum, K. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco: Holden-Day, Inc.

Chemical Rubber Company (CRC). (1970) *Standard Mathematical Tables.* Selby, S., ed. 18th edition. Cleveland, OH.

Clement International Corporation. (1990a) Health effects and dose-response assessment for hydrogen chloride following short-term exposure. Unpublished report prepared for EPA Office of Air Quality Planning and Standards.

Clement International Corporation. (1990b) Health effects and dose-response assessment for acrolein following short-term exposure. Unpublished report prepared for EPA Office of Air Quality Planning and Standards.

Cox, D.; Lindley, D. (1974) *Theoretical Statistics*. London: Chapman & Hall.

Crump, K. (1984) A new method for determining allowable daily intakes. *Fund. Appl. Toxicol.* 4:854-871.

Crump, K. (1985) Mechanisms leading to dose-response models. In: Ricci P., ed. *Principles of Health Risk Assessment.* Englewood Cliffs, NJ: Prentice Hall. pp. 321-372.

Crump, K.; Hoel, D.; Langley, H.; Peto, R. (1976) Fundamental carcinogenic processes and their implications to low dose risk assessment. *Cancer Res.* 36:2973-2979.

Crump, K.; Howe, R. (1985) A review of methods for calculating confidence limits in low dose extrapolation. In: Krewski, D., ed. *Toxicological Risk Assessment.* Canada: CRC Press, Inc.

Dourson, M.; Stara, J. (1983) Regulatory history and experimental support for uncertainty (safety) factors. *Reg. Toxicol. Pharmacol.* 3:224-238.

Dourson, M.; Hertzberg, R.; Hartung, R.; Blackburn, K. (1985) Novel methods for the estimation of acceptable daily intake. *Toxicol. Ind. Health* 1:23-41.

Gaylor, D. (1989) Quantitative risk analysis for quantal reproductive and developmental effects. *Environ. Health Perspect.* 79:243-246.

Gaylor, D.; Slikker, W., Jr. (1990) Risk assessment for neurotoxic effects. *NeuroToxicol.* 11:211-218.

Haseman, J. (1984) Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. *Environ. Health Perspect.* 58:385-392.

Hattis, D.; Lewis, S. (1992) Reducing uncertainty with adjustment factors. *The Toxicologist.* 12(1):1327.

Howe, R.; Crump, K. (1982) GLOBAL 82: A computer program to extrapolate quantal animal toxicity data to low doses. Prepared for the Office of Carcinogen Standards, OSHA, U.S. Department of Labor, Contract 41USC252C3.

Johnson, K.; Gorzinski, S.; Bodner, K.; Campbell, R.; Wolf, C.; Friedman, M.; Mast, R. (1986) Chronic toxicity and oncogenicity study on acrylamide incorporated in the drinking water of Fischer 344 rats. *Toxicol. Appl. Pharmacol.* 85:154-168.

Katz, G.; Krasavage, W.; Terhaar, C. (1984) Comparative acute and subchronic toxicity of ethylene glycol monopropyl ether and ethylene glycol monopropyl ether acetate. *Environ. Health Perspect.* 57:165-175.

Kendall, M. (1951) *The Advanced Theory of Statistics.* Vol. 1. 5th edition. New York: Hafner Publishing Company

Kimmel, C.; Gaylor, D. (1988) Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Anal.* 8:15-21.

Kodell, R.; Howe, R.; Chen, J.; Gaylor, D. (1991) Mathematical modelling of reproductive and developmental toxic effects for quantitative risk assessment. *Risk Anal.* (to appear).

Kupper, L.; Portier, C.; Hogan, M.; Yamamoto, E. (1986) The impact of litter effects on dose-response modeling in teratology. *Biometrics* 42:85-98.

Lehmann, E. (1975) *Nonparametrics. Statistical Methods Based on Ranks.* San Francisco: Holden-Day, Inc.

Melnick, R. (1984) Toxicities of ethylene glycol and ethylene glycol monoethyl ether in Fischer 344/N rats and B6C3F1 mice. *Environ. Health Perspect.* 57:147-155.

Miller, R.; Ayres, J.; Calhoun, L.; Young, J.; McKenna, M. (1981) Comparative short-term inhalation toxicity of ethylene glycol monomethyl ether and propylene glycol monomethyl ether in rats and mice. *Toxicol. Appl. Pharmacol.* 61:368-377.

National Center for Toxicological Research (NCTR). (1981) Teratological evaluation of sulfamethazine. Prepared for Research Triangle Institute. July 8, 1981. RTI-48/31U-2077.

National Research Council (NRC). (1977) *Drinking Water and Health*. Safe Drinking Water Committee. National Academy of Sciences, Washington, DC.

Office of Science and Technology Policy (OSTP). (1985) Chemical carcinogens; A review of the science and its associated principles. *Fed. Reg.*, Part II, pp. 10371-10442. In: *Risk Analysis: A Guide to Principles and Methods for Analyzing Health and Environmental Risks. Appendix G.* Executive Office of the President, 1989. NTIS PB89-137 772.

Peto, R.; Pike, M.; Day, N.; Gray, R.; Lee, P.; Parish, S.; Peto, J.; Richards, S.; Wahrendorf, J. (1980) Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments. Annex. In: *Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal.* IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Supplement 2. International Agency for Research on Cancer, Lyon. pp. 311-426.

Rai, K.; Van Ryzin, J. (1985) A dose-response model for teratological experiments involving quantal response. *Biometrics* 41:1-9.

Sanders, O.T.; Zepp, R.L.; Kirkpatrick, R.L. (1974) Effect of PCB ingestion on sleeping times, organ weights, food consumption, serum corticosterone and survival of albino mice. *Bull. Environ. Contam. Toxicol.* 12:394-399.

SAS. (1988) *SAS/STAT User's Guide*, Release 6.03 edition. SAS Institute, Inc., Cary, NC.

Steel, R.G.; Torrie, J.H. (1980) *Principles and Procedures of Statistics. A Biometrical Approach.* 2d edition. New York: McGraw-Hill Book Company.

Tarone, R.; Ware, J. (1977) On distribution-free tests for equality of survival distributions. *Biometrika* 64:156-160.

Tukey, J.; Ciminera, J.; Heyse, J. (1985) Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* 41:295-301.

U.S. Environmental Protection Agency (U.S. EPA). (1986) Guidelines for the health assessment of suspect developmental toxicants. *Fed. Reg.* 50:39426-39436.

U.S. Environmental Protection Agency (U.S. EPA). (1987) The risk assessment guidelines of 1986. Office of Health and Environmental Assessment, Washington, DC. EPA/600-8-87-045.

U.S. Environmental Protection Agency (U.S. EPA). (1988a) Proposed guidelines for assessing male reproductive risk. *Fed. Reg.* 53:24850-24969.

U.S. Environmental Protection Agency (U.S. EPA). (1988b) Proposed guidelines for assessing female reproductive risk. *Fed. Reg.* 53:24834-24847.

U.S. Environmental Protection Agency (U.S. EPA). (1989) Risk assessment guidance for Superfund. Vol. I: Human health evaluation manual. Interim final. Office of Emergency and Remedial Response, Washington, DC.

U.S. Environmental Protection Agency (U.S. EPA). (1991) IRIS. Background document (4/1/91). Office of Health and Environmental Assessment, Environmental Criteria and Assessment Office, Cincinnati, OH.

APPENDIX A—STATISTICAL METHODS

A.1. BMD Approach

This section describes the statistical procedures associated with the fitting of the BMD models to experimental data. The likelihood approach to parameter estimation is presented as are the methods used to evaluate the fit of the models to the data.

Maximum Likelihood Procedures for Quantal Endpoints. Consider an experiment with g dose levels $d_1, ..., d_g$, and let N_i and X_i, respectively, be the number of animals tested and the number of animals affected at the ith dose level. Let P(d) be the probability that an animal is affected when exposed to a dose d. Assuming that X_i has a binomial distribution with parameters N_i and P(d), the likelihood of the data can be written as

$$L_Q = \prod_{i=1}^{g} \binom{N_i}{X_i} P(d_i)^{X_i}[1-P(d_i)]^{N_i-X_i}$$

The parameters that define P(d) are the only unknowns; they are estimated by the values that maximize the value of L_Q (Cox and Lindley, 1974).

Maximum Likelihood Procedures for Continuous Endpoints. Consider an experiment with g dose levels $d_1, ..., d_g$, and let N_i be the number of animals in the ith dose group, and let x_{ij}, $j = 1, ..., N_i$, $i = 1, ..., g$ represent the response of the jth animal in the ith dose group. It is assumed that x_{ij} has a normal distribution with mean $m(d_i)$ and variance σ_i^2. The unknown parameters in the model consist of the parameters defining m(d) (see Table 4 of the text), plus $\sigma_1, ..., \sigma_g$. Let x_i be the sample mean in the ith dose group, i.e.,

$$x_i = \Sigma_j x_{ij}/N_i$$

where the sum runs from 1 to N_i. Let s_i^2 be the sample variance for group i, i.e.,

$$s_i^2 = \frac{\Sigma_j (x_{ij}-x_i)^2}{(N_i-1)},$$

where, again, the sum runs from 1 to N_i. Then the likelihood of the data can be written as

$$L_C = (2\pi)^{-n/2}\prod_{i=1}^{g} \sigma_i^{-1}\exp\{-[(N_i-1)s_i^2 + N_i(x_i-m(d_i))^2]/2\sigma_i^2\}.$$

The parameters of the continuous BMD model, as well as the variances $\sigma_1^2, ..., \sigma_g^2$, are estimated by those values that maximize the value of L_C (Cox and Lindley, 1974).

Calculation of Confidence Intervals. The "likelihood method" (Cox and Lindley 1974; Crump and Howe, 1985) is recommended for calculating confidence limits (e.g., lower limits on

dose corresponding to a pre-specified level of risk). For example, for quantal data and one of the quantal BMD models, the BMD corresponding to an extra risk of 0.05 and a 95 percent confidence limit is determined as the smallest d that simultaneously satisfies

$$[P(d)-P(0)] / [1-P(0)] = 0.05$$

and

$$2\log(L_{Q_{max}}/L_Q) = (1.645)^2,$$

for some values of the model parameters, where $L_{Q_{max}}$ is the (fixed) maximum value of the likelihood, L_Q is the likelihood as a function of the model parameters, and 1.645 is the 95th percentile of a standard normal distribution.

To calculate a BMD from continuous data, the same procedure is followed except that an equation incorporating the selected measure of adverse response for continuous data replaces the equation for extra risk.

Goodness-of-Fit Tests. Once the parameters of a BMD model have been estimated in the manner described above, the fit of the model to the observed data can be evaluated. For quantal endpoints, an approximate chi-square test is employed; for continuous endpoints, an F test is performed.

For quantal responses, the observed values are numbers of responders and the models predict numbers of responders. The chi-squared test statistic, C, is

$$C = \sum_i \left[\frac{[X_i - N_i * P(d_i)]^2}{N_i * P(d_i) * [1-P(d_i)]} \right]$$

where the sum runs from 1 to g and the notation here is the same as that presented earlier. The degrees of freedom associated with this test are normally g-[number of parameters estimated]. If some of the parameter estimates fall on the boundary of the parameter space, the degrees of freedom are approximated as follows (Anderson, 1983). From the number of dose groups, subtract 1 for estimating the parameter c (the background rate) and subtract 1 for each of the other parameters for which the maximum likelihood estimate is not a boundary value.[1]

The value of C may be compared to the quantiles of a chi-square distribution. For example, if C equals or exceeds the quantile for $(1-\alpha)$ where $\alpha = 0.01$, then we may conclude that the model did not fit the observed data.

[1]The parameters in the quantal BMD models are constrained to lie within certain ranges (see Table 3). A parameter estimate may equal one of the values that define the range for the parameter, in which case a degree of freedom is not lost.

For continuous responses, the mean squared error for lack of fit is compared to the mean squared error associated with pure error to determine if a continuous model has fit the data. The sum of squares associated with the pure error is

$$SS_e = \Sigma\ (N_i-1)*s_i^2,$$

which has $df_e = \Sigma(N_i-1)$ degrees of freedom. In both cases the sum runs from 1 to g and N_i and s_i^2 are as defined above. The sum of squares associated with lack of fit is

$$SS_f = \Sigma\ N_i*(\bar{x}_i-m(d_i))^2,$$

which has df_f degrees of freedom. The value of df_f is equal to the number of dose groups, g, minus 1 (for the estimation of the background parameter c) minus 1 for each of the other parameters for which the estimated value is not equal to a boundary value.

The test statistic

$$F' = [SS_f/df_f] / [SS_e/df_e]$$

is distributed according to an F distribution with degrees of freedom df_f and df_e. The value of F' can be compared to tabulated quantiles of the F distribution with the specified degrees of freedom (Bickel and Doksum, 1977; CRC, 1970) to determine if the model fits the data. For example, when F' equals or exceeds the quantile corresponding to $(1-\alpha)$, where $\alpha = 0.01$, then we may conclude that the model did not fit the observed data.

Application of the BMD Approach to Two Dose Groups. Although the BMD models listed in Table 4 involve three or more parameters, the recommended method for computing statistical bounds will provide a unique lower bound dose even when the data are for only two dose groups (e.g., a control group and one treatment group). For the QQR and CQR models, the lower bound is the same as the one that would have been obtained had the parameter d_0 been fixed at $d_0 = 0$. For the QW and CP models, the lower bound is the same as the one that would have been obtained if the parameter k had been fixed at k = 1. This value of k makes the models assume a linear, no-threshold form. Similar results apply to other models.

Unlike the statistical bounds, the maximum likelihood estimate (MLE) of dose obtained using the models will not be unique when there are only two dose groups. If an MLE is required in such a situation, it is recommended that it be calculated using the models and constraints discussed in the previous paragraph (i.e., $d_0 = 0$ for the QQR and CQR models and k = 1 for the QW and CP models). These selections will generally provide the lowest possible MLE of dose corresponding to a fixed, small level of increased response.

Computer Programs. The fitting procedures described above require sophisticated optimization routines involving iterative numerical calculations. K.S. Crump Division of Clement International has developed software to perform the calculations and to evaluate the fit of models to the data. The software implements the QQR, QW, CQR, and CP models, among others. That software was used for all the examples discussed in this document.

A.2. Statistical Determination of a NOAEL

A NOAEL is defined as the highest experimental dose at which there is no statistically or biologically significant increases in frequency or severity of adverse health effects, compared to corresponding controls. Thus, there should be no statistically significant evidence of a relationship between dose and response for doses up to the NOAEL. Although pairwise tests that compare a single treatment group to the control group are generally used in determining NOAELs, trend tests are available that make use of the data from all of the dose groups up to and including the putative NOAEL. These procedures test for the presence of a trend toward increased responses at increasingly higher doses. These tests incorporate more of the data than pairwise tests; consequently they are generally more powerful.

NOSTASOT Dose. Tukey et al. (1985) have proposed a procedure for determining a no statistical significance of trend (NOSTASOT) dose. This procedure has greater power for determining dose relationships than do multiple pairwise tests (Tukey et al., 1985) and can be used to define a NOAEL. The procedure is described as follows.

First, select a suitable trend test. The selection of such a test depends on the type of endpoint in question and the data available for analysis. Recommended tests for the situations likely to arise in the analysis of non-cancer health effects are presented below.

After selecting the appropriate trend test, apply the test to all of the dose groups. If the test indicates no significant trend, then the highest dose may be considered to be a NOAEL.[2] If the test applied to all the dose groups detects a significant trend, then the highest dose group cannot be a NOAEL. In that case, delete the highest dose group from consideration and repeat the trend test. The highest dose level for which there is no statistically significant trend is the NOAEL (NOSTASOT dose), if biological/ toxicological considerations do not suggest otherwise.

Recommended Trend Tests. Trend tests are proposed here for continuous endpoints and quantal endpoints.

For quantal endpoints, the Mantel-Haenszel trend test (Haseman 1984) is recommended. The Mantel-Haenszel trend test relies on the following test statistic:
where $E_i = N_i^*(\Sigma X_i/\Sigma N_i)$, d_i is the dose level for group i, N_i is the number of animals tested in group i, X_i is the number of animals with the endpoint of interest in group i, and

[2]Some judgment may be required because in certain circumstances the absence of a significant trend when considering all the doses may reflect biological realities that cannot be accounted for by a single trend test. As an example, consider an experiment with a compound that causes two effects. Suppose the occurrence of one of the endpoints makes the observation of the second endpoint less likely (e.g., death or resorption in developmental toxicity studies obscures the occurrence of malformations). In such an instance, the lack of significant trend for the second endpoint, when considering all the dose groups, may reflect the fact that the first endpoint is occurring so often in the high-dose group(s) that the second endpoint cannot be detected in as many animals and consequently makes the trend for that endpoint nonsignificant.

$$Z = \frac{\sum d_i(X_i - E_i)}{V^{1/2}},$$

$$V = \frac{(\sum N_i - \sum X_i) * (\sum X_i) * [(\sum N_i) * (\sum N_i * d_i^2) - (\sum N_i * d_i)^2]}{(\sum N_i)^2 * (\sum N_i - 1)}.$$

In all these equations the summations run over all dose groups. The significance of the Mantel-Haenszel test can be determined by comparing the value of Z with quantiles from a standard normal table (Bickel and Doksum, 1977; CRC, 1970). At the 5 percent level of significance, for example, $Z \geq 1.645$ indicates a significant trend.

The Mantel-Haenszel test as stated may not be appropriate whenever there are significant differences in survival. An important case is one in which the presence of the toxic effect is only identified at necropsy and it is not a fatal effect (i.e., does not cause the death of the animal). In this case the period of observation for the experiment can be divided into subintervals within which there is relatively little variation in death times. The X_i, N_i, E_i, and V values can be calculated as described above for each subinterval. A new Z statistic is calculated as

$$\frac{\Sigma_k \Sigma_i d_i(X_{ik} - E_{ik})}{(\Sigma_k V_k)^{1/2}},$$

where X_{ik} is the number of animals with the toxic effect among animals in the ith treatment group that die in the kth subinterval, E_{ik} is the corresponding expected number based on animals that die in the kth subinterval, and V_k is the corresponding variance. The significance of this test is also evaluated by comparing this statistic with quantities from a standard normal table.

Modifications of the Mantel-Haenszel test that are appropriate when either (1) the toxic effect causes the death of the animal, or (2) the effect can be identified prior to the death of the animal, are discussed in Peto et al. (1980).

For continuous endpoints, Jonckheere's trend test is recommended (Lehmann, 1975). This is a nonparametric test that is an extension of the Mann-Whitney (Wilcoxon) test. A nonparametric test is recommended here because with such tests one need not make assumptions about the distribution of the endpoint under consideration. Given the variety of endpoints that may be analyzed under the NOAEL approach, the lack of distributional assumptions with the Jonckheere test may be advantageous.

To apply Jonckheere's test, one must have the individual observations (i.e., the values of the endpoint for each animal examined). When working from summary reports of the experiments (especially those found in the published literature), these individual values may not be available. In such a case, we recommend the following likelihood ratio trend test based on the CP model.

First, fit the CP model to the data with the power, k, set equal to one. Second, apply the CP model with the dose coefficient, q_1, set equal to zero and k still equal to one. In each run, the log-likelihood is maximized; let us denote the values of the two log-likelihoods as LL_1 and LL_2 for the first run and the second run, respectively. Then,

$$CHI = 2*(LL_1-LL_2)$$

is a likelihood ratio test statistic that is distributed approximately as a chi-square with one degree of freedom under the null hypothesis of no treatment effect. The statistic CHI tests whether the linear dose coefficient is significant (i.e., whether a significant dose-related trend exists). Comparison of CHI with the one-degree-of-freedom chi-squared quantile corresponding to $(1-2\alpha)$ determines whether the trend is significant for significance level α, based on a one-sided (directional) test of trend.

Alternatively, versions of nonparametric trend tests that are extensions of log-rank and Wilcoxon tests (Tarone and Ware, 1977) may be applied to either quantal or continuous data.

Pairwise Tests. As alternatives to the trend tests listed above, one may wish to employ pairwise tests to determine if a dose group is significantly different from the control group irrespective of the overall trend. As noted, however, the trend tests have greater power for detecting a significant dose-related increase than do the pairwise tests (Tukey et al., 1985). The problem of multiple comparisons must also be considered when doing many pairwise tests. Nevertheless, pairwise tests may provide useful supplementary information that can be used in addition to the NOSTASOT approach.

For quantal data, Fisher's exact test is the recommended pairwise test (Bickel and Doksum, 1977). For continuous data, a nonparametric approach is recommended for the pairwise comparisons as well as the trend tests. The Mann-Whitney (Wilcoxon) test is suitable in cases where the individual data are available (Lehmann, 1975). When group means and standard deviations are available but the individual results are not available, t-tests may be applied to test for pairwise differences (Bickel and Doksum, 1977). The nonparametric approach is preferred when the individual data are available because it avoids distributional assumptions.

Computer Programs. Statistical software packages such as SAS (SAS, 1988) contain programs that can implement most of the statistical tests discussed for the NOSTASOT procedure.

GLOSSARY

Adverse effect. A biochemical change, functional impairment, or pathological lesion that either singly or in combination adversely affects the performance of the whole organism or reduces an organism's ability to respond to an additional environmental challenge.

Benchmark dose (BMD). A statistical lower confidence limit on the dose producing a predetermined, altered response for an effect.

Benchmark response (BMR). A predetermined level of altered response or risk at which the benchmark dose is calculated.

Biologically significant effect. A response in an organism or other biological system that is considered to have a substantial or noteworthy effect (positive or negative) on the well-being of the biological system. Used to distinguish statistically significant effects or changes, which may or may not be meaningful to the general state of health of the system.

Cancer. A malignant growth.

Carcinogenic. Able to produce malignant tumor growth. Operationally, most benign tumors are usually included also.

Chronic exposure. Long-term exposure usually lasting six months to a lifetime.

Confidence limit. A confidence interval for a parameter is a range of values that has a specified probability (e.g., 95 percent) of containing the parameter. The confidence limit refers to the upper or lower value of the range (e.g., upper confidence limit).

Continuous endpoint. A measure of effect that is expressed on a continuous scale (e.g., body weight or serum enzyme levels).

Critical effect. The first adverse effect, or its known precursor, that occurs as the dose rate increases.

Critical study. A bioassay performed on the most sensitive species used as the basis of RfD determination.

Developmental toxicity. Adverse effects on the developing organism that may result from exposure prior to conception or postnatally to the time of sexual maturation. Adverse developmental effects may be detected at any point in the life span of the organism. Major manifestations of developmental toxicity include death of the developing organism; induction of structural abnormalities (teratogenicity); altered growth; and functional deficiency.

Dose-response relationship. A relationship between (1) the dose, either "administered dose" (i.e., exposure) or absorbed dose, and (2) the extent of toxic injury produced by that chemical. Response can be expressed either as the severity of injury or proportion of exposed subjects affected. A dose response assessment is one of the four steps in a risk assessment.

Endpoint. An observable or measurable biological or chemical event used as an index of the effect of a chemical on a cell, tissue, organ, organism, etc.

Extrapolation. An estimate of response or quantity at a point outside the range of the experimental data. Also refers to the estimation of a measured response in a different species or by a different route than that used in the experimental study of interest (i.e., species-to-species, route-to-route, acute-to-chronic, high-to-low).

Genotoxic. A broad term that usually refers to a chemical that has the ability to damage DNA or the chromosomes. This can be determined directly by measuring mutations or chromosome abnormalities or indirectly by measuring DNA repair, sister-chromatid exchange, etc. Mutagenicity is a subset of genotoxicity.

Lifetime. Covering the life span of an organism (generally considered 70 years for humans).

Lowest observed adverse effect level (LOAEL). The lowest dose or exposure level of a chemical in a study at which there is a statistically or biologically significant increase in the frequency or severity of an adverse effect in the exposed population as compared with an appropriate, unexposed control group.

Maximum likelihood estimate (MLE). A statistical best estimate of the value of a parameter from a given data set.

Model. A mathematical representation of a natural system intended to mimic the behavior of the real system, allowing description of empirical data and predictions about untested states of the system.

Neurotoxicity. Ability to damage nervous tissue.

No observed adverse effect level (NOAEL). An exposure level at which there are no statistically or biologically significant increases in the frequency or severity of adverse effects between the exposed population and its appropriate control; some effects may be produced at this level, but they are not considered as adverse, nor precursors to adverse effects. In an experiment with several NOAELs, the regulatory focus is primarily on the highest one, leading to the common usage of the term NOAEL as the highest exposure without adverse effect.

Pharmacokinetics. The field of study concerned with defining, through measurement or modeling, the absorption, distribution, metabolism, and excretion of drugs or chemicals in a biological system as a function of time.

Population variability. The concept of differences in susceptibility of individuals within a population to toxicants due to variations such as genetic differences in metabolism and response of biological tissue to chemicals.

Quantal endpoint. A dichotomous measure of effect; each animal is scored "normal" or "affected" and the measure of effect is the proportion of scored animals that is affected.

Reference concentration (RfC). An estimate (with uncertainty spanning perhaps an order of magnitude) of a continuous inhalation exposure to the human population (including sensitive

subgroups) that is likely to be without an appreciable risk of deleterious non-cancer effects during a lifetime.

Reference dose (RfD). An estimate (with uncertainty spanning perhaps an order of magnitude) of a daily exposure to the human population (including sensitive subgroups) that is likely to be without appreciable risk of deleterious non-cancer effects during a lifetime.

Reproductive toxicity. Harmful effects on fertility, gestation, or offspring caused by exposure of either parent to a substance.

Risk. The probability of injury, disease, or death under specific circumstances, relative to the background probability. In quantitative terms, risk is expressed in values ranging from zero (representing the certainty that the probability of harm is no greater than the background probability) to one (representing the certainty that harm will occur).

Risk assessment. The scientific activity of evaluating the toxic properties of a chemical and the conditions of human exposure to it in order both to ascertain the likelihood that exposed humans will be adversely affected and to characterize the nature of the effects they may experience. The assessment may involve some or all of the following four steps:

Hazard identification. The determination of whether a particular chemical is or is not causally linked to particular health effect(s).

Dose-response assessment. The determination of the relation between the magnitude of exposure and the probability of occurrence of the health effects in question.

Exposure assessment. The determination of the extent of human exposure.

Risk characterization. The description of the nature and often the magnitude of human risk, including attendant uncertainty.

Spontaneous. Arising in the absence of external causes.

Statistically significant effect. In statistical analysis of data, a health effect that exhibits differences between a study population and a control group that are unlikely to have arisen by chance alone.

Subchronic exposure. Exposure to a substance spanning no more than approximately 10 percent of the lifetime of an organism.

Threshold toxicant. A substance showing an apparent level of effect that is a minimally effective dose, above which a response may occur and which dose no response is expected.

Uncertainty. In the conduct of risk assessment (hazard identification, dose-response assessment, exposure assessment, risk characterization) the need to make assumptions or best judgments in the absence of precise scientific data creates uncertainties. These uncertainties, expressed qualitatively and sometimes quantitatively, attempt to define the usefulness of a particular evaluation in making a decision based upon the available data.

Uncertainty factor (UF). One of several, generally 10-fold factors, used in operationally deriving the reference dose (RfD) from experimental data. UFs are intended to account for (1) the variation in sensitivity among the members of the human population; (2) the uncertainty in extrapolating animal data to the case of humans; (3) the uncertainty in extrapolating from data obtained in a study that is of less-than-lifetime exposure; and (4) the uncertainty in using LOAEL data rather than NOAEL data.